Universidade do Minho
Escola de Ciências

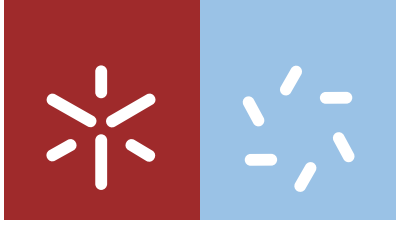
Andreia Alves Forte de Oliveira Monteiro

Contributions to Spatial and Temporal Modelling

janeiro de 2019

Andreia Alves Forte de Oliveira Monteiro **Contributions to Spatial and Temporal Modelling**

UMinho | 2019



Universidade do Minho
Escola de Ciências

Andreia Alves Forte de Oliveira Monteiro

Contributions to Spatial and Temporal Modelling

Tese de Doutoramento em Matematica Aplicada das Universidades do Minho, Aveiro e Porto, MAP-PDMA

Trabalho efetuado sob a orientação da
Professora Doutora Raquel Menezes
e da
Professora Doutora Maria Eduarda Silva

janeiro de 2019

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.

Universidade do Minho, 30 / 01 / 2019

Full name: : Andreia Alves Fonte de Oliveira Monteiro

Signature: : Andreia Alves Fonte de Oliveira Monteiro

Acknowledgements

Firstly, I'd like to thank my supervisors, Professor Raquel Menezes and Professor Maria Eduarda Silva. Complete this work would have been impossible without their constant support, availability and constructive suggestions.

I also want to thank my colleague and friend Ana for her encouraging words during these years.

My acknowledgments must be addressed to the Portuguese Foundation for Science and Technology (FCT) for funding this research through the Individual Scholarship PhD PD/BD/ 105743/2014, and also as a member of the research project PTDC/MAT-STA/28243/2017. One word of acknowledgment to Minho University for providing me a rich work environment and to the Center for Research & Development in Mathematics and Applications of Aveiro University (CIDMA) for financial support within project UID/MAT/04106/2019.

Last but not the least, I would like to thank my family, especially André and Mariana for their permanent support in this academic adventure.

To my father memory.

To my family

Abstract

Contributions to Spatial and Temporal Modelling

Recent technological advances allow the collection of data in space and time in a wide range of contexts such as environmental and health sciences. Most of these data are generated by monitoring processes and present spatial and/or temporal structures. Traditionally spatial and temporal modelling assumes that the locations (in time or space) sampled are either fixed or stochastically independent of the spatial and temporally continuous phenomenon under study. However, it is well-known that, for example, in air pollution studies, typically the monitors are placed near the most likely pollution sources in areas of high population density. In context of medical studies, a patient is usually observed most frequently when he presents a worse clinical condition. In these examples neither are the observations obtained regularly in time/space nor are the observed locations (in time or space) stochastically independent of the phenomenon under study. Ignoring this dependence can lead to biased estimates and misleading inferences. In this work, we consider the problem of modelling time series with informative observation times. We introduce the concept of Preferential Sampling in the temporal dimension and we discuss alternative model-based approaches to make inference and prediction under stochastic sampling schemes. In the first approach, we present a model to deal with irregularly spaced time series in which the sampling design depends on the contemporaneous value of the underlying process, under the assumption of a Gaussian response variable. For this model, we present two estimation methods, one based on Monte Carlo simulations and the other based on a Laplace approximation. The second approach proposes a model for irregularly spaced time series in

which the sampling design depends on all past history of the observed processes. All discussed model-based approaches are illustrated with numerical studies.

Keywords: Preferential Sampling, time series, continuous time autoregressive process, SPDE, evolutionary processes.

Resumo

Contribuições para a Modelação Espacial e Temporal

Os recentes avanços tecnológicos permitem a recolha de dados no espaço e no tempo numa grande variedade de contextos, como nas ciências ambientais e da saúde. A maior parte desses dados é gerada por processos de monitorização e apresenta estruturas espaciais e/ou temporais. Tradicionalmente, a modelação espacial e temporal assume que as localizações amostradas (no tempo ou no espaço) são fixas ou estocasticamente independentes do fenómeno espacial e temporal em estudo. No entanto, é bem conhecido que, por exemplo, em estudos de poluição do ar, normalmente as estações de monitorização são colocadas perto das fontes de poluição mais prováveis em áreas de alta densidade populacional. Em estudos médicos, um paciente é geralmente observado com maior frequência quando apresenta pior condição clínica. Nestes exemplos, nem as observações são obtidas de forma regular no tempo/espaço, nem as localizações das observações (no tempo ou no espao) são estocasticamente independentes do processo em estudo. Ignorar essa dependência pode levar a estimativas tendenciosas e inferências enganosas. Neste trabalho, consideramos o problema de modelar séries temporais com tempos de observação informativos. Introduzimos o conceito de Amostragem Preferencial na dimensão temporal e discutimos diferentes abordagens baseadas em modelos para fazer inferência e previsão debaixo deste esquema de amostragem. Numa primeira abordagem, apresentamos um modelo para lidar com séries temporais irregularmente espaçadas em que o desenho amostral depende do valor contemporâneo do processo subjacente, sob a hipótese de uma variável de resposta Gaussiana.

Para este modelo, apresentamos dois métodos de estimação, um baseado em simulações de Monte Carlo e outro baseado numa aproximação de Laplace. Na segunda abordagem, propomos um modelo para séries temporais irregularmente espaçadas, nas quais o desenho amostral depende de toda a história passada dos processos observados. Os modelos propostos são ilustrados com estudos numéricos.

Palavras-Chave: Amostragem Preferencial, séries temporais, processos autoregressivos contínuos no tempo, equações diferenciais parciais estocásticas, processos evolucionários.

Contents

Glossary of Terms

APE	Absolute Prediction Error
AR	Autoregressive
BLUE	Best Linear Unbiased Estimator
CAR	Continuous Time Autoregressive
NO_2	Nitrogen Dioxide
MLE's	Maximum Likelihood Estimates
MCMLE	Monte Carlo Maximum Likelihood Estimates
PS	Preferential Sampling
SPDE	Stochastic Partial Differential Equation
GMRF	Gaussian Markov Random Field
LAP	Laplace
TMB	Template Model Builder
GF	Gaussian Field
EVOL	Evolutionary

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

1.1 Motivation

In recent years data indexed in space and time have become the norm rather than the exception as a result from technological developments. As an example we may mention that sensors and mobile devices routinely gather data. Summarizing, modelling and inferencing from these multidimensional and usually large data sets present new challenges to statistical science. Most of these data present spatial and/or temporal dependence structures that require new methodological and computational tools. The literature on the analysis of multidimensional data, in particular space-time indexed data, is increasingly abundant but many open problems remain.

To motivate one such problem, consider environmental monitoring data obtained at a set of conveniently located sites over time, so common nowadays. Often the temporal dimension of the data is characterized by high resolution and multiple seasonal patterns while the sites are predominantly located in cities. Modelling these spatio-temporal data commonly assumes that the sampled locations (in time or space) are either fixed or stochastically independent of the phenomenon under study. However, it is well-known that in air pollution studies the monitors are, typically, placed near the most likely pollution sources and in areas of high population density. Thus, the aforementioned assumption fails (in the spatial dimension) since the process under study determines the data-locations. This problem, coined Preferential Sampling in the context of spatial statistics has been discussed in a model-based approach by Diggle *et al.* (2010). Similarly, data may be prone to irregular spacing in time for various

reasons. For example, data related to natural disasters such as earthquakes, floods, or volcanic eruptions which typically occur at irregular time intervals, give rise to irregularly or unevenly spaced time series. A particular situation of irregularly spaced data is that in which the sampling design depends also, for practical constraints, on the observed values. Examples occur in fisheries where the data are observed when the resource is available, in sensoring when sensors keep only some records in order to save memory, in clinical studies, when a worse clinical condition leads to more frequent observations of the patient and, in a completely different scenario, the times at which transactions occur in the financial markets depend largely on the value of the underlying asset. In all such situations, there is stochastic dependence between the process under study and the times at which the observations are made, and the observation times are informative on the underlying process. This thesis aims at contributing to jointly model time series with informative observation times.

1.2 Main Objectives

Our framework considers joint models for data indexed by informative observation times, assuming a continuous time underlying process observed at regular or at irregular and stochastic points. Under a regular sampling, the choice of adequate covariates or time functions prove to be enough to deal with informative times. Under unevenly spaced time series, more complex model-based approaches are discussed in this thesis. To represent the underlying process we opt for a continuous time series model such as the Continuous Time Autoregressive (CAR) model, which is mathematically and computationally tractable and yet sufficiently flexible to represent a wide range of phenomena. The assumption that the observation times are informative and stochastic is equivalent to assuming that they are a realisation of a random process, which is stochastically dependent on the underlying process. This dependence is specified via two model-based approaches. The first extends the concept of Preferential Sampling to the temporal dimension, allowing the dependence between the underlying process and the sampling (observation) process to be contemporaneous only. In the second model-based approach, the dependence encompasses also the past history of the processes (underlying and observation) which is more realistic in many real life contexts. Both approaches rely on point processes: a log Cox Gaussian approach for the Preferential

Sampling scenario and marked evolutionary processes to include the history of the process.

Thus, the major objectives of this work are:

- to introduce the concept of Preferential Sampling in the temporal dimension
- to develop model-based approaches that take into account the stochastic dependence of the sampling design on the process of interest
- to develop a computationally feasible framework to make inference and prediction.

We consider maximum likelihood estimation of the model parameters. Since the likelihood involves an unobserved process, the underlying process, we resort to simulation and numerical techniques to achieve its minimization. In a first approach, we consider Monte Carlo simulation which proves to have a prohibitive computational cost. In a second approach that overcomes the computational problems of parameter estimation, we use an alternative numerical method based on the Laplace approximation of the marginal likelihood and adopt a technique based on stochastic partial differential equation (SPDE) to approximate the CAR process.

1.3 Thesis outline

In Chapter 2, we introduce main concepts related to traditional methods that may support the analysis of data under irregular sampling in space and in time. The usage of the variogram as a tool to measure spatial dependence between samples is highlighted. The consideration of a convenient continuous time domain dynamic model for the underlying continuous time stationary process, such as the Continuous time AR (CAR) process, to deal with irregularly spaced time series, is also highlighted.

In Chapter 3, a two-step approach is suggested to model the spatial and temporal dynamics of spatio-temporal data sets characterized by irregular sampling locations and high resolution in the temporal dimension. The approach is applied to the data set comprising hourly measurements of NO_2 at 49 stations located over Portugal, being the informative times treated by an harmonic regression and adequate covariates.

In Chapter 4, we introduce the concept of Preferential Sampling in the temporal dimension as a formal definition for the dependence between the process generating

the times of the observations and the data values. We propose a framework to deal with Preferential Sampling in time, also able to deal with irregularly spaced time series, under Preferential Sampling or not. We proceed with likelihood inference to estimate the parameters of this model. We first consider a Monte Carlo approach for maximum likelihood estimation of the model and then we consider a numerical method based on a Laplace approach to optimize the likelihood. Numerical studies with simulated and real data sets are performed to illustrate the benefits of this model based approach versus the traditional one which ignores Preferential Sampling issues.

In Chapter 5, we consider that the sampling design may depend on all past history of the process and we propose a model, based on evolutionary processes that takes into account that the times and values of the observations contain important information for the underlying process (informative and stochastic time points). Using numerical studies, we document the performance of this approach comparing the results of shared parameter estimates with those obtained from the traditional approach for irregularly spaced data.

In Chapter 6, we present some conclusions and directions for future work.

2

Dependent data under irregular sampling

In this Chapter, our main objective is to present a concise review of main concepts and methods traditionally adopted for the analysis of data collected under irregular sampling in space and in time. For irregular sampling in space, we introduce the main fields of spatial statistics, that may support the analysis of this type of data, namely Geostatistics. For irregular sampling in time (our focus in this work), a literature review of current methods is given and we introduce the continuous time autoregressive processes (CAR).

2.1 Introduction

Recent technological advances allow the collection of data in space and time in a wide range of contexts such as environmental and health sciences. Data can present a spatial structure, determined by the locations where data are collected, and/or a temporal one, determined by the frequency with which observations are taken at these locations. It is acknowledge that data collected in space, like data collected over time, tend to exhibit statistical dependence. One commonly exhibited form of dependence is spatial continuity, which reflects the fact that observations taken at two sites tend to be more alike if the sites are close together than if the sites are far apart. Examination of this correlation/dependence in time is commonly referred to as time series analysis. The existence of spatial or temporal dependence in the data invalidates the results of a clas-

sical statistical analysis, based on an assumption of independent observations, involved.

Spatial models work with data collected from different spatial locations. Over the past 20 years, spatial statistics have emerged. One factor that has contributed to this rise is the tremendous increase in computational capability.

Spatial data can be thought of as resulting from observations on the stochastic process

$$\{Y(\mathbf{x}) : \mathbf{x} \in D\} \tag{2.1}$$

where $\mathbf{x}$ are locations within some spatial region $D \subset \mathbb{R}^d$, typically, $d = 1, 2$ or 3.

Cressie (1993) states that it is not reasonable to assume that spatial locations of data occur regularly. The choice of locations is commonly guided by external factors, by the context of the investigation or practical issues. For example, in air pollution studies, the monitors are typically placed near the most likely pollution sources in areas of high population density. In other applications, the choice of sample points may be restricted in some way. One form of restriction is when the study region includes sub-regions which are of interest for prediction but inaccessible for sampling.

On the other hand, analysis of experimental data that have been observed at different points in time leads to specific problems in statistical modelling and inference. The correlation introduced by the sampling of points in time can limit the applicability of conventional statistical methods. Equally spaced sampling, is perhaps, the most frequently assumed sampling scheme in practice. However irregularly spaced or unevenly spaced time series occur in many situations, for example in natural disasters like volcanic eruptions and earthquakes, in economics, climatology and environment sciences. Another example occurs in observational astronomy, measurements of properties such as the spectra of celestial objects are taken at irregularly spaced times determined by seasonal, weather conditions, and availability of observation time slots. In clinical studies (or more generally, longitudinal studies), a patient's state of health may be observed only at irregular time intervals, and different patients are usually observed at different points in time.

2.2 Irregular sampling in space

To analyse spatial data, according to Cressie (1993), we can identify three major spatial processes, namely lattice processes, point processes and continuous processes commonly referred to as geostatistics.

Lattice data

In this type of data D in (2.1) is a countable collection of spatial sites, with well defined boundaries, at which data are observed. The collection D of such sites is called a lattice, which is then supplemented by neighborhood information. The hypothesis underlying modelling lattice data is that adjacent regions share information in the sense that close areas have more in common than distant areas. The usual model structure in these is the conditional autoregressive model, Besag (1991). These models induce autoregressive spatial autocorrelation through an adjacency structure of the lattice units.

Lattice data include, for example, pixel values from remote sensing of natural resources, presence or absence of a plant species in square blocks laid out over a prairie remnant and the number of deaths of a cancer type in the counties of a nation, or other administrative districts. Typically, this type of data is irregularly spaced but not stochastic.

Point processes

The aim of point processes is to analyse the geometrical structure of patterns formed by objects that are distributed randomly in space. Examples include locations of trees in a forest stand, blood particles on a glass plate and galaxies in the universe. In addition to the location of these objects, there may be further variables that are of interest associated with each point. This information is known as marks. In this situation, objects are represented by points and marks resulting in a marked point process. The points describe the locations of the objects, and the marks provide additional information, thus characterizing the objects further, e.g. through their type, size or shape. This type of data, typically, is irregularly spaced and stochastic.

In other words, the objective of point process statistics is to understand and describe the short-range interaction among the points and explain the mutual positions

of the points. Quite often this concerns the degree of clustering or repulsion (inhibition) among points and the spatial scale at which these operate. The analysis of a point pattern also provides information on underlying processes that have caused the patterns, Illian *et al.* (2008). Within the spatial point processes, perhaps the most important theoretical development over the last years has been the provision of formal, likelihood-based methods of inference for a reasonably wide range of models, Diggle (2013).

There is an extensive literature on point processes, ranging from rather theoretical to more applied texts, for e.g. Daley & Vere-Jones (2003, 2008), Møller & Waagepetersen (2004), Illian *et al.* (2008) and Diggle (2013).

Geostatistics

The term geostatistics identifies the part of spatial statistics which is concerned with data obtained by spatially discrete sampling of a spatially continuous process. Originally, the term geostatistics was coined by Georges Matheron and co-workers in France, to describe their work dealing with problems of spatial prediction resulting from mining industry. However, the geostatistical methods are now used in many areas of application, far beyond the mining context in which they were originally developed. Common examples are meteorological and air pollution data. In this context, the main goals are to determine a spatial pattern, modelling correlation/covariance, make predictions and testing whether there exists a spatial structure or not. For a model-based approach see Diggle & Ribeiro (2007).

Following the notation used by Diggle & Ribeiro (2007), we denote a set of geostatistical data by $(\mathbf{x_i}, y_i) : i = 1, ..., n$, where $\mathbf{x_1}, ..., \mathbf{x_n}$ are locations within an observation region $D \subset \mathbb{R}^2$ and $y_1, ...y_n$ are measurements associated with these locations. It is assumed the existence of an unobserved field process $\{S(\mathbf{x}) : \mathbf{x} \in D\}$, usually regarded as our goal of prediction. In many applications it is also assumed that:

- S is a stationary [1] and isotropic [2] Gaussian process, with mean μ, variance σ^2 and spatial correlation function $\rho(u) = \text{Corr}\{S(\mathbf{x}), S(\mathbf{x}')\}$, where $u = \|\mathbf{x} - \mathbf{x}'\|$

[1] A random process is second-order stationary if its first moment is a constant and the covariance between two variables is a function of the difference between their locations.

[2] A random process is isotropic if it remains invariant when subject to rotation of coordinates.

and $\|\cdot\|$ denotes the Euclidean distance;

- conditional on S, the y_i are realizations of mutually independent random variables $Y_i = Y(\mathbf{x_i})$, normally distributed with conditional means $E[Y_i|S(\cdot)] = S(\mathbf{x_i})$ and conditional variances τ^2.

- $Y(\mathbf{x_i}) = S(\mathbf{x_i}) + N(0, \tau^2), i = 1, ..., n.$

According to Cressie (1993), the variogram is a model-based measure of the spatial statistical dependence in a geostatistical process. The variogram, $2\gamma(\cdot)$ or the semivariogram, $\gamma(\cdot)$ of a spatial stochastic process $S(\mathbf{x})$ is the function

$$\gamma(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\text{Var}\left\{S(\mathbf{x}) - S(\mathbf{x}')\right\}$$

Note that, $\gamma(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\left[\text{Var}\left\{S(\mathbf{x})\right\} + \text{Var}\left\{S(\mathbf{x}')\right\}\right] - 2\text{Cov}\left\{S(\mathbf{x}), S(\mathbf{x}')\right\}$. In the stationary and isotropic case, this simplifies to $\gamma(u) = \sigma^2\left\{1 - \rho(u)\right\}$, which explains the inclusion of the one-half factor in the definition of the variogram.

Because the mean of a stationary process is constant, the variogram in the stationary case can also be defined as $\gamma(u) = \frac{1}{2}\text{E}\left[\left\{S(\mathbf{x}) - S(\mathbf{x} - u)\right\}^2\right]$. Now, assume that the data is generated by the stationary process

$$Y(\mathbf{x_i}) = S(\mathbf{x_i}) + N(0, \tau^2), i = 1, ..., n$$

Then the variogram of the observation process $\gamma_Y(u)$ is defined by $\gamma_Y(u_{ij}) = \frac{1}{2}\text{E}\left[(Y_i - Y_j)^2\right]$, where $u_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|$. It follows that

$$\gamma_Y(u) = \tau^2 + \sigma^2\left\{1 - \rho(u)\right\} \tag{2.2}$$

Typically, $\rho(u)$ is a monotone decreasing function and equation (2.2) neatly summarizes the essential qualities of a classical geostatistical model. The typical variogram is a monotone increasing function with the following features. The intercept, τ^2, corresponds to the nugget variance, which occurs as a result of small variability between spatially correlated variables and/or measurement errors. The asymptote, $\tau^2 + \sigma^2$, corresponds to the variance of the observation process Y, sometimes called the sill, which in turn is the sum of the nugget variance and the signal variance, σ^2. The way in which the variogram increases is determined by the correlation function $\rho(u)$. When $\rho(u) = 0$ for u greater than some finite value, this value is known as the range of the variogram.

The range is the distance such that pairs of spatial locations further than this distance apart are negligibly correlated. A geostatistical convention defines the practical range as the distance u_0 at which $\rho(u_0) = 0.05$

One of the most critical properties characterizing a variogram is that of *conditional negative-definiteness*, i.e. the requirement that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{x_i} - \mathbf{x_j}) \leq 0$$

for any finite set of spatial locations, $\{\mathbf{x_1}, ...\mathbf{x_n}\}$, and for any set of real numbers $\{a_1, ...a_n\}$, such that $\sum_{i=1}^{n} a_i = 0$.

In the absence of this property, the mean square prediction error could be estimated by an absurd negative value. This leads to the impossibility of using some variogram estimators within the inference and prediction context. One possible approach to solve this problem is to approximate the empirical variogram (*variogram based on observation or experiment*) by any parametric model which is known to be valid. The idea is to search, among the families of valid variograms, for one that best approximates the underlying spatial dependence of the available sample data.

As example of an useful correlation function adopted in geostatistical data modelling, we have the Matérn family with algebraic form given by

$$\rho(u) = \left(2^{\nu-1}\Gamma(\nu)\right)^{-1} \left(\frac{u}{\phi}\right)^{\nu} K_\nu \left(\frac{u}{\phi}\right)$$

where $\nu > 0$ and $\phi > 0$ are parameters, and $K_\nu(\cdot)$ denotes a Bessel function of order ν. The parameter ϕ, the range, determines the rate at which the correlation decays to zero with increasing u. The parameter ν determines the analytic smoothness of $S(\mathbf{x})$. For $\nu = 0.5$ the Matérn correlation function reduces to the exponential, $\rho(u) = \exp(-u/\phi)$, whilst $\nu \to +\infty$, $\rho(u) = \exp(-(u/\phi)^2)$ which is called the Gaussian correlation function.

Having identified the model for spatial dependence, we can proceed with predicting the spatially continuous process at an unsampled location. The process of spatial prediction 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 since 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** estimator.

In Geostatistics, we can have regular or irregular sampling but, the usual assumption is that the selection of the sampling locations does not depend on the values of the spatial variable. In fact, most techniques are based on the assumption, possibly tacit, that sampling locations are uniformly distributed over the observed region. However, there are situations in which the process under study determines the data-locations and the above mentioned assumption is violated. Diggle *et al.* (2010) coined this phenomenon as Preferential Sampling: the sampling process and the observed process are dependent and there is an underlying stochastic relationship between data and locations.

2.3 Irregular sampling in time

Real time series sometimes exhibit various types of "irregularities": missing observations, observations collected not regularly over time for practical reasons, observation times driven by the series itself, or outlying observations. However, the vast majority of methods of time series analysis are designed for regular time series only. There are few methods available in the literature for the analysis of irregularly spaced series. Some authors, such us Jones (1981, 1985), Belcher *et al.* (1994) and Brockwell (2009) have suggested an embedding into continuous diffusion processes, with the aim of using the well established tools for univariate autoregressive moving average (ARMA) processes, as opposed to the development of a complete set of tools for equally spaced data.

It must be noted that sometimes equally spaced time series are treated as irregularly spaced time series, namely time series with missing observations and multivariate data sets that consist of time series with different frequencies, even if the observations of each time series are reported at regular intervals. One of the first authors to treat evenly sampled gene expression time series with missing values as unevenly sampled data is Ruf (1999).

Observations with irregularly spaced sampling times are much harder to work with, partly because the established and efficient algorithms developed for equally spaced sampling times are no longer applicable, Li (2014). A common approach to perform parametric estimation is to construct a log-likelihood function in terms of the unknown

parameters, Brockwell (2001). When the sampling times are considered deterministic, the traditional approach is to build the classical Gaussian log-likelihood function. However, because the inversion of the covariance matrix has to be performed, numerical evaluation of this Gaussian log-likelihood function is in general very expensive, Lange (2010). This computational effort may be overcome regulating the sampling scheme, with some form of interpolation and then considering it as being equally spaced. Under the assumption of equally spaced sampling times, the Gaussian log-likelihood function can be approximated, at least for a sufficiently large sample, by the Whittle log-likelihood function, Whittle (1961). This approach has been successfully applied to irregularity caused by missing values, Little & Rubin (2014). While, it may be reasonable to use this methodology to deal with the minor irregularities in sampling times caused by missing values, the interpolation procedure will typically change the dynamic of the underlying process, leading to biased estimates for the parameters, Erdogan *et al.* (2005). Moreover, there is little understanding of which particular interpolation method is the most appropriate on a given data set.

Another approach is to consider, a convenient continuous time dynamic model for the underlying continuous time stationary process such as the Continuous time ARMA (CARMA) model. The application of Kalman recursion techniques to the parametric estimation of CARMA processes is reviewed in Tómasson (2015). Additionally, Kelly *et al.* (2014) estimate the parameters of an irregularly sampled CARMA process using a Bayesian framework.

Although there are different approaches to deal with irregularly spaced time series, in this work, we will concentrate on the analogues of Autoregressive (AR) model in discrete time but in continuous time, the CAR model, namely the first order CAR process. Whereas the AR(p) model is a difference equation, CAR(p) models are defined by stochastic differential equations, as is to be expected when generalizing to continuous time. The relevant statistical theory of CAR(1) process is reviewed in the next Section and much of the required theory is given by Jones (1981) and Priestley (1981).

CAR(1) process

A stochastic process $X = \{X(t) : t \geq 0\}$ is a continuous time autoregressive process of order p, CAR(p), if satisfies the differential equation:

$$X^{(p)}(t) + \alpha_{p-1}X^{(p-1)}(t) + \alpha_{p-2}X^{(p-2)}(t) + ... + \alpha_0 X(t) = dW(t) \tag{2.3}$$

where, $\alpha_0, ...\alpha_{p-1}$ are constants, $\{W(t) : t \geq 0\}$ is a Wiener process with variance parameter σ_w^2 and $X^{(i)}(t)$ is the i-th derivative of $X(t)$, $dW(t)$ is interpreted as the increments of $W(t)$ in the time interval $(t, t + dt)$.

The simplest CAR process is the CAR(1) satisfying:

$$X^{(1)}(t) + \alpha_0 X(t) = dW(t)$$

which can be written in differential form as:

$$dX(t) + \alpha_0 X(t)dt = dW(t) \tag{2.4}$$

where $\alpha_0 > 0$ is the autoregressive coefficient. Note that, $X(t)$ is asymptotically stationary if an only if $\alpha_0 > 0$.

Under these conditions, equation (2.4) has a unique stationary solution, Hyndman (1992):

$$X(t) = X(0)e^{-\alpha_0 t} + e^{-\alpha_0 t}\int_0^t e^{\alpha_0 u}dW(u),\ t \geq 0$$

Expected value

If $X(0) = C$

The expected value of $X(t)$ is defined as

$$\begin{aligned} E[X(t)] &= E\left[X(0)e^{-\alpha_0 t} + e^{-\alpha_0 t}\int_0^t e^{\alpha_0 u}dW(u)\right] \\ &= Ce^{-\alpha_0 t} + e^{-\alpha_0 t}E\left[\int_0^t e^{\alpha_0 u}dW(u)\right] \\ &= Ce^{-\alpha_0 t} \end{aligned} \tag{2.5}$$

Since $\alpha_0 > 0$ we have

$$\lim_{t\to\infty} E[X(t)] = 0$$

Variance and Covariance

The variance of $X(t)$ is defined as

$$\begin{aligned}
\text{Var}[X(t)] &= E[(X(t) - E[X(t)])^2] \\
&= E[(e^{-\alpha_0 t}\int_0^t e^{\alpha_0 u} dW(u))^2] \\
&= E\left[e^{-2\alpha_0 t}\left(\int_0^t e^{\alpha_0 u} dW(u)\right)^2\right] \\
&= e^{-2\alpha_0 t}\sigma_w^2 \int_0^t e^{2\alpha_0 u} du \\
&= e^{-2\alpha_0 t}\sigma_w^2 \left(\frac{e^{2\alpha_0 t}}{2\alpha_0} - \frac{1}{2\alpha_0}\right) \\
&= \frac{\sigma_w^2}{2\alpha_0}\left(1 - e^{-2\alpha_0 t}\right) \qquad (2.6)
\end{aligned}$$

Note that,

$$\lim_{t\to\infty} \text{Var}[X(t)] = \frac{\sigma_w^2}{2\alpha_0}$$

Let $C(h) = \text{Cov}[X(t), X(t+h)]$ denote the covariance function of the CAR(p) process $X(t)$.

The characteristic equation

$$A(z) = \sum_{j=0}^{p} \alpha_j z^j = 0 \qquad (2.7)$$

with $\alpha_p = 1$ has q distinct roots $\lambda_1, ..., \lambda_q$ where λ_i has multiplicity m_i.

Using contour integration, (Doob, 1953, p.543) showed that

$$C(h) = \sigma_w^2 \sum_{i=1}^{q} c_i(h) e^{\lambda_i |h|} \qquad (2.8)$$

where $c_i(h)$ is a polynomial in h of order m_i. Where all the roots are distinct ($m_i = 1, \forall i$), Jones (1981) gives

$$c_i(h) = \left[-2Re(\lambda_i) \prod_{\substack{l=1 \\ l \neq i}}^{p} (\lambda_l - \lambda_i)(\overline{\lambda}_l + \lambda_i) \right]^{-1}$$

where $Re(\lambda_i)$ is the real part of λ_i, and $\overline{\lambda}_l$ denotes the complex conjugate of λ_l.

If $p = 1$ then (2.7) reduces to

$$A(z) = \alpha_0 + z = 0 \Rightarrow \lambda_1 = -\alpha_0$$

and (2.8) to

$$C(h) = \frac{\sigma_w^2}{2\alpha_0} e^{-\alpha_0 |h|} \tag{2.9}$$

It follows that the autocorrelation function, $\rho(h)$ of $X(t)$ is

$$\rho(h) = \frac{C(h)}{C(0)} = e^{-\alpha_0 |h|} \tag{2.10}$$

We would like to point out that for one dimensional position vector, the exponential spatial correlation structure is equivalent to the CAR(1) structure. In particular, as explained in Pinheiro & Bates (2001), if one considers $\alpha_0 = \frac{1}{\phi}$, then ϕ is the correlation parameter, generally referred as range in spatial statistics. Additionally, CARMA processes driven by Wiener process are Gaussian processes.

2.4 Summary

In this Chapter, we present a concise review of the main methodologies used to analyse data under irregular sampling in space and time, defining some important concepts of

Geostatistics and CAR processes, serving as baseline for the modelling approaches to be discussed in this thesis.

Firstly, in the next Chapter we present a methodology for data sets irregular in space but regular in time, characterized by high resolution in the temporal dimension and multiple seasonal patterns.

3

Modelling data irregular in space and regular in time: a case study

Part of the work included in this Chapter was published in Monteiro *et al.* (2017). The objective is to model the spatial and temporal dynamics of spatio-temporal data sets characterized by irregular sampling locations and high resolution in the temporal dimension, which are becoming the norm rather than the exception in many application areas, namely environmental modelling. A two-stage modelling approach is proposed, which combined with a block bootstrap procedure correctly assesses uncertainty in parameters estimates and produces reliable confidence regions for the space-time phenomenon under study.

3.1 Introduction

It is acknowledged that air pollution is a social as well as an environmental problem, leading to a multitude of adverse effects on human health, ecosystems and the built environment. Several research studies, systematic reviews and meta-analysis have been carried out to analyse health effects of air pollutants: Shin *et al.* (2008) considered these issues by monitoring the risk of death associated with outdoor air pollution; McCarthy *et al.* (2009) used ambient monitoring data to determine the relative importance of individual air toxics for chronic cancer and noncancer exposures; Lai *et al.* (2013) analysed the risk estimates for mortality and morbidity outcomes due to air pollutants; Keramatinia *et al.* (2016) studied the relationship between exposure to NO_2 and

breast cancer incidence; Song *et al.* (2016) conducted a systematic review to provide an association between air pollution and cardiac arrhythmia. In fact, the European Environment Agency, EEA (2015) considers air pollution the single largest environmental health risk in Europe. Thus the need for accurate assessment of air pollution arises not only to investigate the linkage between ambient exposure and health effects but also with regard to compliance with legislated regulatory standards to control levels of environmental exposure. The above considerations advance the need for statistical models aimed at characterizing and predicting air quality events and assessing policies over specified areas.

In Portugal, estimation of the index of air quality involves measurements of the following chemical elements: carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), ozone (O_3) and fine particulate matter as PM_{10}. The index is based on the pollutant with the highest concentration relative to the Portuguese annual limit values for the protection of human health. This work focus on NO_2 concentrations, which is considered a primary pollutant, formed naturally in the atmosphere by lightning and produced by plants, soil and water Carslaw (2005). However, the major sources are the fossil fuel combustion processes, the emissions from electricity generating stations and road traffic. Furthermore, NO_2 concentration levels closely follow vehicle emissions, in many situations, thus providing a reasonable marker exposure to traffic. Nitrogen dioxide is toxic by inhalation and there is evidence that long-term exposure to NO_2 at high concentrations has adverse health effects, namely in respiratory and cardiovascular systems, Ricciardolo *et al.* (2004). NO_2 and other nitrogen oxides are also precursor of ozone and particulate matter, whose effects on human health and the environment are well documented. Concentrations of NO_2 have been analysed extensively in many urban areas (Carslaw, 2005; Grice *et al.*, 2009; Roberts-Semple *et al.*, 2012) as well as in background sites (Donnelly *et al.*, 2011; Menezes *et al.*, 2016). Moreover, these studies acknowledge that meteorological conditions influence NO_2 levels (Donnelly *et al.*, 2011; Russo & Soares, 2014; Shi & Harrison, 1997). Thus the overall results indicate recurrent multiple seasonal patterns resulting from anthropogenic activity and the influence of meteorological variables. Fassò & Negri (2002) propose a non-linear statistical model to deal with the problem of high frequency and multiple periodicities underlying environmental data dynamics. De Livera *et al.* (2011) also consider complex seasonal

patterns into their modelling approaches, using exponential smoothing. The former works restrict their applications to one geographical location.

This work proposes a methodology to characterize the spatial and high resolution temporal evolution of spatio-temporal data using geostatistical approaches. The approach takes into account that environmental data often incorporate distinct recurring patterns in time and considers the influence of meteorological variables. The suggested framework is applied to hourly NO_2 concentration levels in Portugal. Spatio-temporal statistical modelling aims at revealing dependencies and spatio-temporal dynamics, e.g. Cameletti *et al.* (2011) and, in our particular case, at obtaining hourly concentration predictions over the country. To this end the model proposed by Menezes *et al.* (2016) is extended to hourly data and meteorological variables are included. A block bootstrap procedure is proposed to correctly assess uncertainty of parameters estimates, as well as to produce reliable confidence regions for (space-time) NO_2 concentrations. The model is potentially useful in many areas including assessment of environmental impact and environmental policies.

3.2 The Portuguese data set

This study analyses hourly measurements of NO_2 obtained from the online database on air quality (Qualar, 2015) of the Portuguese Environment Agency, whose mission is to propose, develop and monitor the public policies for the environment and sustainable development. The database on air quality provides hourly measurements, resulting from monitoring activities, for various pollutants, including NO_2. The available data include information about the type of site where the station is placed (background, industrial or traffic) and the environment of the zone (urban, suburban or rural). The most serious drawback of QualAr is that validated data are only made available in October of the following year.

The hourly NO_2 concentrations under analysis concern 49 stations located over Portugal (mainland) from October 1st to December 31st in 2014, in a total of 108192 observations. From the 49 stations, 33 are classified as background, 10 as traffic and 6 as industrial, 29 are located in urban areas, 11 in rural areas and 9 in suburban areas, Figure 3.1. The selected period corresponds to the highest NO_2 levels along the year,

according Menezes *et al.* (2016), who analysed daily NO_2 data during 8 years. This study has about 18% of missing data in the hourly levels of NO_2.

The NO_2 concentrations have a mean of 20.6 $\mu g/m^3$, standard deviation of 21.9 and median of 13$\mu g/m^3$. The histogram of NO_2 concentrations, represented in Figure 3.2, reveals asymmetry indicating departure from Gaussianity.

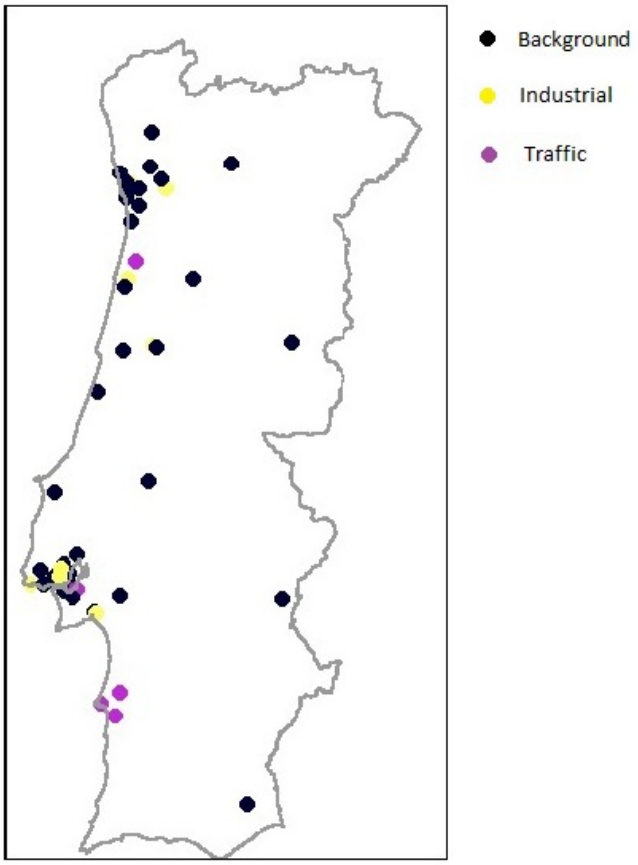


Figure 3.1: Monitoring Network.

A periodogram analysis of the data reveals periodicities at 12, 24 and 168 hrs, which corresponds to intra-daily, daily and weekly periods. These recurring patterns are clearly observed in Figure 3.3, which represents mean hourly values for both weekdays and weekends. NO_2 levels show two daily peaks, one in the morning (8:00) and one in the afternoon (18:00) which coincide with rush-hour traffic, with the second peak

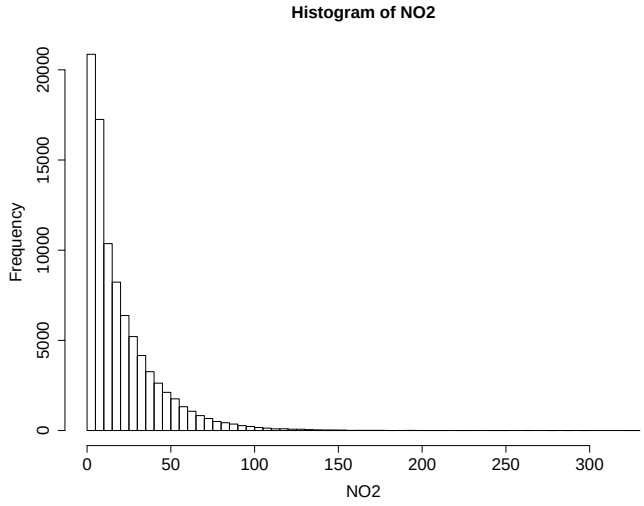


Figure 3.2: Histogram of NO_2 concentrations.

being more pronounced than the first. Moreover, the mean NO_2 concentrations are much lower on weekends (particularly on Sunday) than on weekdays, displaying, also, smaller variation on weekends, which reflect reduced levels of vehicular emissions on non-working days. Thus, the two main seasonal effects in the data: intra-day as well as intra-week periodicities, may be, at least partially, explained by characteristics of the station. In fact, Figure 3.4 illustrates the influence of the location and the environment of the station in values of NO_2. It is clear that the stations located in traffic areas and urban zones present higher values for their NO_2 quartiles as well higher variability. This analysis indicates that the type of site and the environment zone must be considered as explanatory variables.

Since it is acknowledge that meteorological variables influence NO_2 levels, hourly data from the following meteorological variables were obtained from Weather Underground (2015), which provides weather data collected hourly from around the world: wind speed (km/h); air temperature (^{0}C) and relative humidity (%). The analysis of the correlation between these meteorological variables and NO_2 levels identified the well known negative associations among them. High NO_2 concentrations are favored by cold and drier weather; on the other hand, an increase of wind-speed, generally, promotes dilution and dissipation of the pollutants, thus yielding lower levels of NO_2, in accor-

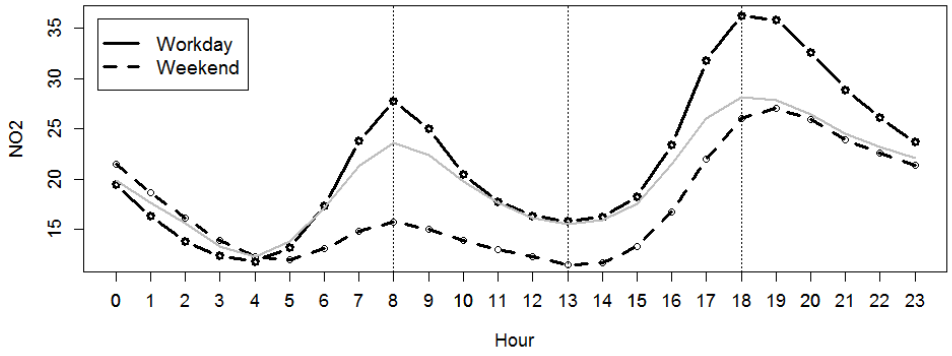


Figure 3.3: Mean NO_2 concentrations, for workdays and weekends. The gray line identifies the trigonometric representation of the cyclical component based on Fourier series, Section 3.3.

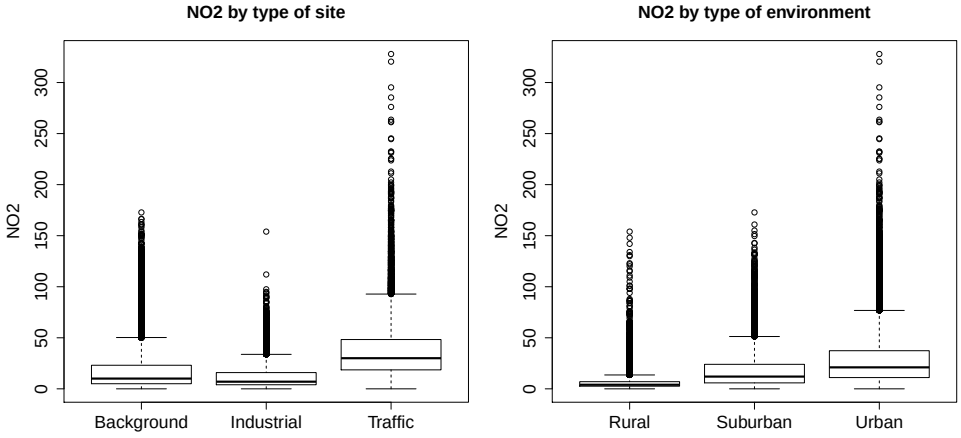


Figure 3.4: Boxplots of NO_2 concentrations, by type of site and type of environment.

dance with Shi & Harrison (1997). Additionally Spearmans rank correlation coefficient between NO_2 and the meteorological variables for several lags, represented in Figure 3.5 indicates that the strongest correlations occur at 6-hour lag with air temperature, 1-hour lag with wind-speed and 5-hour lag with relative humidity. Therefore, these meteorological variables at the identified lags are considered as explanatory variables

for NO_2 levels.

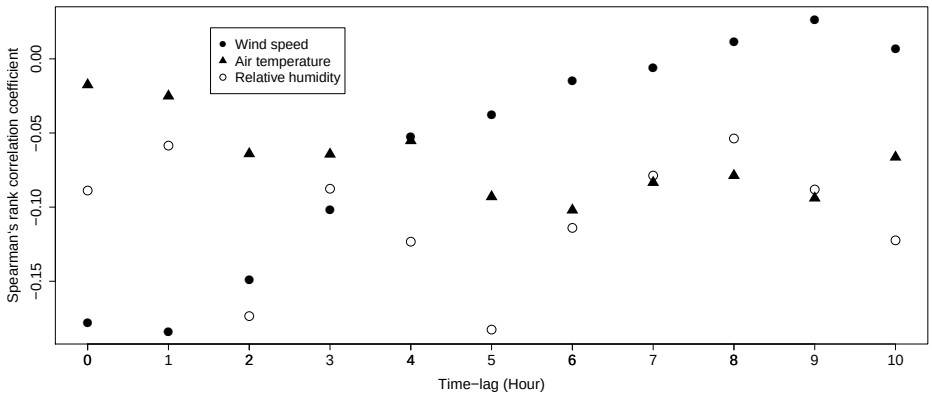


Figure 3.5: Spearmans rank correlation coefficient between hourly NO_2 and the meteorological variables for several lags.

The above exploratory analysis indicates two main seasonal effects in the temporal dynamics of NO_2 levels: daily and weekly. This preliminary study also shows that the variables type of site (background, industrial or traffic) and type of environment (urban, suburban or rural), together with the meteorological variables air temperature (6-hour lag), wind speed (1-hour lag) and relative humidity (5-hour lag) are possible explanatory variables for the NO_2 levels. Further analysis, not reported here, shows the presence of strong spatial dependence in the NO_2 data set as widely reported in environmental pollution data literature.

These remarks evidence the importance of using a spatio-temporal model incorporating multiple seasonalities for describing the complex structure and dynamics of the phenomenon.

3.3 Methodology

Consider a spatio-temporal stochastic process $Y(\mathbf{s}, t)$ indexed in space by $s \in \mathbb{R}^d$ and in time by $t \in \mathbb{N}$. The process can be represented as

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \delta(\mathbf{s}, t) \tag{3.1}$$

where $\mu(\mathbf{s},t) = \mathrm{E}(Y(\mathbf{s},t))$ represents a spatio-temporal mean field modelling the trend, usually refered to as the large-scale variation component and $\delta(\mathbf{s},t)$ is a zero-mean smooth stationary spatio-temporal process that models the small-scale variation (hereafter referred to as stationary residual).

3.3.1 Large-scale variation

The mean component $\mu(\cdot)$ in the above model may be a deterministic function when the physics of the underlying phenomenon is known. However, in the large majority of problems and spatio-temporal data sets such knowledge is unavailable and we must resort to stochastic specifications which aim at representing the patterns of the observed variability. Accordingly, in the specification of the mean component we include regression variables observed jointly with the response variables and incorporate, also, complex nested or non nested seasonal and cyclic effects. In fact, many time series exhibit multiple seasonal patterns: hourly pollution levels reveal a daily pattern with period of 12 or 24 as well as a weekly pattern with period $24 \times 7 = 168$ and a long series might also exhibit an annual seasonal pattern with period 24×365, resulting from the natural cycles and anthropogenic activity. Thus, a flexible approach to model (3.1) consists on considering the generalized linear model (GLM) which combines three components:

- A random component specifying the conditional distribution of the response variable $Y(\mathbf{s},t)$, given the values of explanatory variables. This conditional distribution may be any from the exponential family thus avoiding transformations of the response variable.

- A systematic component which specifies a linear predictor that is a function of a set explanatory variables $\mathbf{X}$

$$\eta(\mathbf{s},t) = \mathbf{AX} \tag{3.2}$$

  where $\mathbf{A}$ is a matrix of real coefficients and $\mathbf{X}$ a matrix of regressors.

  A smooth and invertible linearising link function $g(\cdot)$ which transforms the expectation of the response variable $\mathrm{E}(Y(\mathbf{s},t)) = \mu(\mathbf{s},t)$ into the linear predictor $\eta(\mathbf{s},t) = g(\mu(\mathbf{s},t))$.

Matrix $\mathbf{X}$ contains the K regression variables $X_i(\mathbf{s},t)$, $i=1,\ldots,K$ observed jointly with the response $Y(\mathbf{s},t)$, and the periodic regressors that capture the periodicities in the time series. Assume that there are L identified periods $(m_1,\ldots,m_L)$ and assume for each cyclic component at time t, $S_{t,l}$ a trigonometric representation based on Fourier series with the form $S_{t,l}=\sum_{j=1}^{k_l}\left[\phi_{j,1}\cos\left(\frac{2\pi jt}{m_l}\right)+\phi_{j,2}\sin\left(\frac{2\pi jt}{m_l}\right)\right]$, where k_l represents the number of harmonics required for the lth cyclic component. The number of periodic regressors L depends on the data under study and may be determined by frequency analysis of the time series. Thus we can write

$$\begin{aligned}\eta(\mathbf{s},t) &= \alpha+\sum_{i=1}^{K}\beta_i X_i(\mathbf{s},t)+\sum_{l=1}^{L}S_{t,l}\\ &= \alpha+\sum_{i=1}^{K}\beta_i X_i(\mathbf{s},t)+\sum_{l=1}^{L}\sum_{j=1}^{k_l}\left[\phi_{j,1}\cos\left(\frac{2\pi jt}{m_l}\right)+\phi_{j,2}\sin\left(\frac{2\pi jt}{m_l}\right)\right]\end{aligned} \tag{3.3}$$

where α, β_i, $\phi_{j,1}$, $\phi_{j,2}\in\mathbb{R}$ are regression parameters.

3.3.2 Small-scale variation

It is now necessary to consider the space-time dependence structure underlying the stationary spatio-temporal residual $\delta(\mathbf{s},t)$. Many methods have been proposed in the literature to define valid models for the spatio-temporal dependence structures e.g. (De Cesare *et al.*, 2001; Gneiting, 2002). For a comparative review of the characteristics of many of these currently accepted and implemented models see De Iaco (2010). One of the main distinctions between these models is based on the notion of separability. A separable space-time covariance function can be written as the product of a purely spatial component and a purely temporal component. This allows for efficient estimation (especially computationally), and inference but the separability is restrictive and often require unrealistic assumptions (Bruno *et al.*, 2003), and a major disadvantage of these models is that they can not incorporate the space-time interaction. Thus, in our study, the attention has shifted to non-separable covariance structures, namely the product-sum and sum-metric models, which are widely used in the literature. Other parametric families of non-separable models are discussed in Cressie & Huang (1999), Ma (2008) and Rodrigues & Diggle (2010). For more general classes of non-separable covariance functions see Fonseca & Steel (2011), Ip & Li (2015).

The product-sum model can be defined in terms of the semivariogram as

$$\gamma_{\mathbf{s}t}\left(\mathbf{h}_s, h_t\right) = \gamma_{\mathbf{s}}\left(\mathbf{h}_s\right) + \gamma_t\left(h_t\right) - k\gamma_{\mathbf{s}}\left(\mathbf{h}_s\right)\gamma_t\left(h_t\right) \tag{3.4}$$

where γ_s and γ_t are the corresponding valid semivariogram functions in space and time, $(\mathbf{h}_s, h_t) \in \mathbb{R}$ and

$$k = \frac{sill_s + sill_t - sill_{st}}{sill_s\, sill_t}$$

where $sill_s$ and $sill_t$ represent the sill of the marginal semivariograms in space and time, repectively, and $sill_{st}$ is the global sill.
The sum-metric model can be defined:

$$\gamma_{\mathbf{s}t}\left(\mathbf{h}_s, h_t\right) = \gamma_s\left(\mathbf{h}_s\right) + \gamma_t\left(h_t\right) + \gamma\left(\left|\mathbf{h}_s\right| + \alpha\left|h_t\right|\right) \tag{3.5}$$

with $\alpha \in \mathbb{R}$ and γ_s e γ_t the semivariograms.

3.3.3 Parameter estimation and inference by block bootstrap

The estimation of model (3.1) is accomplished in a 2-step approach which estimates separately the trend (large-scale variation) and the spatio-temporal dependence structure (small-scale variation) components. First obtain point estimates for the regression parameters using maximum likelihood (ML) and relaxing the assumption of non-correlated errors, underlying ML estimation in GLM. Then fit a valid non-separable space-time variogram to the residuals resulting from the previous step, fully accomplishing the estimation of the spatio-temporal correlation in the data.

An important issue arising in the first step as a consequence of relaxing the assumption of uncorrelated residuals is that of assessing the statistical significance of the estimated parameters. To handle this issue we resort a bootstrap procedure for serially correlated data. We consider a modification of the so called block bootstrap Kreiss & Paparoditis (2011), based on moving and overlapping blocks in the time dimension, when taking fixed data in the space dimension. The main idea consists of dividing the temporal data, $(X_1, \ldots, X_T)$ say, into blocks of consecutive observations of length l, $(X_t, \ldots, X_{t+l-1})$. The first block corresponds to $(X_1, \ldots, X_l)$ and each new block slides M time units, becoming $(X_{1+k\times M}, \ldots, X_{l+k\times M})$ with $k = 1, \ldots, K$, $M << l$ and $l + K \times M \leq T$, allowing for a total of $K+1$ blocks. This bootstrap approach is particularly appropriate when one has long time series, as it is usually the case with hourly

data, collected at a small number of geographical locations. This bootstrap approach further allows to obtain a confidence band for large-scale variation predictions.

The estimation of the parameters in the semivariograms (3.4) and (3.5) relies in a least-squares approach over a space-time empirical variogram. At this stage the sample marginal variograms in space and time, defined in De Iaco & Posa (2012) are important to give some guidance for the selection of the one-dimensional variogram components in (3.4) and (3.5). In fact, the selection of adequate models in (3.4) and (3.5) is crucial to guarantee that the resulting function is valid for prediction using kriging tools. Myers (2004) provide some guidelines that may be useful for model selection. To evaluate the final variograms, a cross-validation approach originally introduced in Stone (1974), and meanwhile adapted to the context of dependent data, is used. This procedure consists on eliminating one observation from the whole set and then predicting its value from the remaining data through a kriging methodology. Repeating the procedure for all the observations, the Mean Square Error (MSE) of the resulting errors can be used to choose between several (variogram) models. Following an adequate choice of a spatio-temporal variogram, a block bootstrap procedure is once more resorted to correctly assess uncertainty in its parameters estimates.

3.4 Results

The preliminary data analysis of NO_2 concentrations in Portugal carried out in Section 3.2, indicates that the underlying process presents several characteristics such as non Gaussianity, multiple periodicities and spatial dependence, for which model (3.1) introduced in Section 3.3 may be particularly useful. The 2-step estimation procedure proposed is carried out leading to the characterization of the mean or large-scale variation component in Section 3.4.1, and that of the stationary residual or small-scale variation component in Section 3.4.2. The estimation procedure is implemented in R environment R Core Team (2015) and the following packages are used: *gstat* Pebesma (2004), *sp* and *space-time* Bivand *et al.* (2013).

3.4.1 Large-scale variation

Firstly, we model the trend of NO_2 data using a Generalized Linear Model as given in equation (3.3). In the case of NO_2 concentrations, exploratory analysis revealed that it

is a continuous variable with an asymmetric distribution, in particular, we assume that the response variable is gamma distributed with log-link. As the gamma distribution is only defined for strictly positive values, we make a translation of the data set by 0.0001. We consider six explanatory variables: type of site, type of environment, if weekend, air temperature (6-hour lag effect), wind speed (1-hour lag effect) and relative humidity (5-hour lag effect). Other factors were also considered, like the distinction between the days of the weekend (week, Saturday and Sunday), but this did not result in significant improvements. Furthermore, we consider other hour lag effects for meteorological variables, however, the best model is selected under Akaike information criterion and by graphical observation of NO_2 fitted values *vs.* NO_2 levels.

For modelling the seasonal effects in the data set, we proceed as represented in (3.3), assuming a trigonometric representation for each cyclic component. The dominant frequencies of the data were estimated, based on those stations without missing values, which made it possible to identify two important periodicities equal to 12 and 24 hours. Consequently, although we have tested distinct periodic regressors, including one for the weekly cycles, the simpler model restricted to the daily (or half-daily) cycles proved to be preferable.

The results of the gamma regression of the hourly NO_2 concentrations are summarized in Table 3.1. The standard errors were obtained using a moving block bootstrap in the time dimension, each block with 5 weeks sliding 3 hours, generating 456 replicates. All 49 monitoring stations were kept as fixed. According to the notation presented in Section 3.3, the block length $l = 5 \times 7 \times 24 = 840$ hours, $\delta = 3$ hours and $K = 455$. Two weeks blocks were also considered, however, these were not able to capture patterns of intra- and inter-day variability, meaning that the seasonal components became no significant in the trend model. From the results in Table 3.1, we conclude that the values of NO_2 concentrations are greater during the week and in monitoring stations where the environment is urban or suburban and the type of site is traffic. Besides that, NO_2 levels increase by a factor of 3.64 from rural to urban, by a factor of 1.64 from background to traffic, and by a factor of 1.22 during the week. In respect of meteorological variables, these variables have significant negative associations with NO_2 levels. These conclusions confirm the results from the preliminary data analysis. Wind speed has a stronger influence on NO_2 concentrations than humidity and air temperature. Furthermore, NO_2 level decrease 3% by an increase of 1 km/h in wind speed

and decrease 1% by an increase of 1% in humidity. In the case of air temperature, the lack of significance of its coefficient was confirmed under the proposed block bootstrap approach, which can be explained by the fact that only months with low temperature are selected (in October to December, mean is 14.6^0C and standard deviation is 5.4^0C). The acquired coefficient of determination shows that 41% of the large-scale variation of NO_2 concentrations is explained under this trend model.

Parameter	Estimate	Over-optimistic Std. Error$^{(*)}$	Bootstrap Std. Error
Intercept	2.452	0.019	0.259
Type of site (baseline: Background)			
Industrial	-0.517	0.009	0.026
Traffic	0.489	0.008	0.031
Day of the week (baseline: Weekend)			
Week	0.202	0.007	0.024
Environment (baseline: Rural)			
Suburban	1.147	0.010	0.128
Urban	1.310	0.008	0.153
Air Temperature	-0.008	0.0006	0.015
Wind Speed	-0.029	0.0004	0.002
Relative Humidity	-0.006	0.0002	0.002
$\sin(\frac{2\pi t}{12})$	-0.228	0.004	0.019
$\cos(\frac{2\times 2\pi t}{12})$	-0.015	0.004	0.007
$\sin(\frac{2\times 2\pi t}{12})$	0.033	0.004	0.011
$\cos(\frac{4\times 2\pi t}{12})$	0.008	0.004	0.002
$\cos(\frac{2\pi t}{24})$	0.093	0.005	0.039
$\sin(\frac{2\pi t}{24})$	-0.167	0.005	0.018
$\cos(\frac{3\times 2\pi t}{24})$	0.018	0.004	0.008
$\sin(\frac{3\times 2\pi t}{24})$	0.102	0.004	0.012
$\cos(\frac{5\times 2\pi t}{24})$	0.016	0.004	0.005
$\sin(\frac{5\times 2\pi t}{24})$	-0.021	0.004	0.004

Table 3.1: Estimates of the gamma regression coefficients for hourly NO_2 concentrations, together with the corresponding standard errors obtained by bootstrap. The standard errors given in (*) were obtained by GLM when relaxing the assumption of non-correlated residuals.

3.4.2 Small-scale variation

Having estimated the large-scale variation $\mu(\mathbf{s},t)$ as $g^{-1}(\eta(\mathbf{s},t))$ in (3.3), we now aim at estimating the dependence structure of the stationary residual $\delta(\mathbf{s},t)$, resulting from $Y(\mathbf{s},t) - \mu(\mathbf{s},t)$ in (3.1). This issue is addressed through the approximation of the spatio-temporal variogram. The fit of the empirical variogram demands estimation of the unknown parameters of the theoretical model, namely, the nugget τ^2, the partial variance σ^2 and the range ϕ. We start by analyzing the marginal spatial and the marginal temporal correlation structures, defined in De Iaco & Posa (2012). The Gaussian model is selected for the approximation of the spatial variogram, suggesting the parameters estimates $\hat{\tau}_s^2 = 0.19$, $\hat{\sigma}_s^2 = 0.59$ and $\hat{\phi}_s = 35.47$km. For the temporal variogram, it is selected the Exponential model, and the resulting parameters estimates are $\hat{\tau}_t^2 = 0.60$, $\hat{\sigma}_t^2 = 0.06$ and $\hat{\phi}_t = 47.47$ hours. We examined other models, however, the results for the parameter estimates were similar.

To decide whether to adopt the product-sum model in (3.4) or sum-metric model in (3.5), we proceed with a cross-validation study to compare both models, according to which the eliminated observations are predicted through the kriging tools. For each model, we estimate the mean error (ME) and the mean square error (MSE) based on all resulting prediction errors. The results in Table 3.2 are very similar, however, the model sum-metric has an extra parameter for anisotropy which allows dealing with spatial and temporal distances in the same term. Besides that, the sum-metric model makes it possible to use specific variogram for space, time, and space-time. Therefore, we decide to choose the sum-metric model with a Exponential function for the temporal component and Gaussian functions for the spatial and the spatio-temporal components.

Model	joint	temporal	space	ME	MSE
Product-sum model	-	Exp	Gau	-0.016	0.214
Sum-metric model	Gau	Exp	Gau	-0.007	0.219

Table 3.2: ME and MSE estimates of the cross-validation study.

Under this selection, the fitted final model is represented in Figure 3.6 (right), being the corresponding empirical variogram given in the left panel. The resulting parameters estimates and corresponding standard errors, obtained by moving block bootstrap,

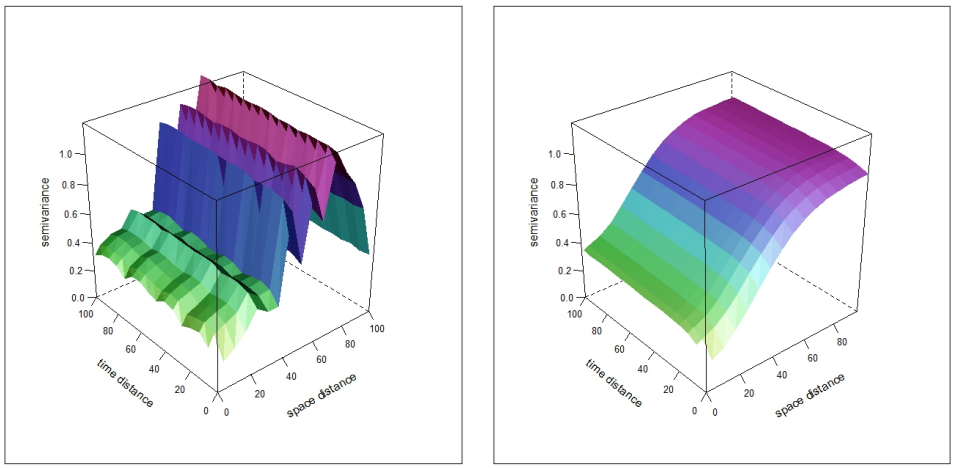


Figure 3.6: Plots of the experimental estimator (left) and the fitted model (right) for the space-time variogram.

Variogram	Model	τ^2	σ^2	ϕ	α
Spatial	Gau	0.015 (0.025)	0.662 (0.128)	40km (1.348)	
Temporal	Exp	0.010 (0.020)	0.071 (0.022)	100h (0.003)	
Joint	Gau	0.172 (0.018)	0.132 (0.030)	70 (0.024)	13.007 (0.074)

Table 3.3: Parameters estimates, and corresponding bootstrap standard errors obtained by moving block bootstrap with blocks of 5 weeks sliding 8 hours, generating 171 replicates, for the spatial, temporal and spatio-temporal variograms.

blocks of 5 weeks sliding 8 hours, generating 171 replicates, are given in Table 3.3. Initially, we tried the option of sliding 3 hours instead of 8 hours, as done for the regression coefficients estimates in the trend, but the computational cost associated to the estimation of the variogram was not acceptable. According to the results, we conclude that the majority of the total variation is explained by the spatial component. The temporal and spatio-temporal components have a smaller contribution. Furthermore, NO_2 concentrations have a significative spatial correlation up to 40 km and a temporal correlation up to 100 hours (approximately 4 days).

3.4.3 Model assessment

To assess the goodness of fit of the model two measures are chosen: the Mean Absolute Percentual Error (MAPE) and the Mean Absolute Scaled Error (MASE). The MAPE,

being a percentage error has the advantage of being scale-independent, and so is frequently used to compare model predictive performance between different data sets, in this case stations with different environment characteristics. On the other hand, the MAPE, being a measure based on percentage errors has the disadvantage of presenting large values for observations close to zero. Hyndman & Koehler (2006) proposed the MASE as an alternative measure based on scaled errors, which, in fact, compare the error in the value predicted by the model with that of a naive prediction. The naive prediction must take into account the data seasonality.

For model assessment the predictions are defined as $\hat{Y}(\mathbf{s},t) = \hat{\mu}(\mathbf{s},t) + \hat{\delta}(\mathbf{s},t)$, where: $\hat{\mu}(\mathbf{s},t)$ is the fitted large scale variation at location $\mathbf{s}$ and time t, given climate conditions; and $\hat{\delta}(\mathbf{s},t)$ is the predicted small scale-variation, obtained under a cross-validation approach. This means that data from station at location $\mathbf{s}$ is eliminated and $\delta(\mathbf{s},t)$ is predicted from the remaining data by kriging tools. Considering T observations for any particular station $\mathbf{s}$, one has

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|e_t|}{Y(\mathbf{s},t)} \times 100\% \qquad \text{MASE} = \sum_{t=1}^{T}\frac{|e_t|}{\sum_{t=1}^{T}|Y(\mathbf{s},t) - Y(\mathbf{s},t-168)|}$$

where $e_t = \hat{Y}(\mathbf{s},t) - Y(\mathbf{s},t)$.

Station	Environment	Type	MASE	MAPE(×100%)
Loures	urban	background	0.841	0.54
Beato	urban	background	0.618	0.40
Entrecampos	urban	traffic	0.898	0.48
Avenida da Liberdade	urban	traffic	0.443	0.47
Matosinhos	suburban	background	0.623	0.51
Lourinhã	rural	background	0.874	0.38
Sonega	rural	industrial	0.757	0.68

Table 3.4: MASE and MAPE errors for some stations, according environment of the zone and type of the site.

Since our data set is of high dimensionality, model assessment is performed for a subset of seven monitoring stations (Loures, Beato, Entrecampos, Avenida da Liberdade, Matosinhos, Lourinhã, Sonega) representative of the different types of environments during five consecutive working days: from 2014-10-13 at 0:00 (Monday) to 2014-10-17

at 24:00 hrs (Friday). Goodness of fit measures, MAPE and MASE for the seven stations are presented in Table 3.4. For the computation of the MASE, a more adequate measure in our case, we considered a naive prediction of the NO_2 concentration at a location, the value of the concentration at that location, at the same day and same time of the day of the previous week, computed for mean climate conditions of that time of day. This procedure takes into account the multiple seasonalities present in NO_2 concentrations. The MASE values range from 0.44 to 0.90 and are all less than one indicating that the model predicts more accurately that the naive predictor. There is not a clear pattern on the errors with urban, traffic stations (Avenida da Liberdade and Entrecampos) presenting the lowest and highest MASE errors. The absence of such a pattern may be explained on one hand by the high variability that hourly concentrations present and on the other hand, the low number of stations classified as rural and industrial.

The predicted large and small scales variation and observed concentration in Loures, an urban and background station, represented in Figure 3.7 illustrates the high variability present in the data. Although the overall mean intra-day pattern of the NO_2 concentrations is well described by the model, see Figure 3.3, individual stations and days present particularities that remain unexplained by the model.

Even so, this assessment exercise allows to conclude that the model provides a good enough representation of the data and can be used for out of sample prediction and scenario generation.

3.5 Space-time prediction and forecasting

This Section illustrates the potential of the proposed spatio-temporal modelling strategy for prediction and forecasting. The former is accomplished by interpolating in the observed space-time dimension, through the kriging tools. The latter is accomplished through the mean predictor given in (3.3), as it allows to obtain NO_2 forecasts as a function of the explanatory variables.

3.5.1 Space-time prediction

A major advantage of the proposed modelling methodology is the possibility of using space-time kriging techniques, namely ordinary kriging, to make predictions at any

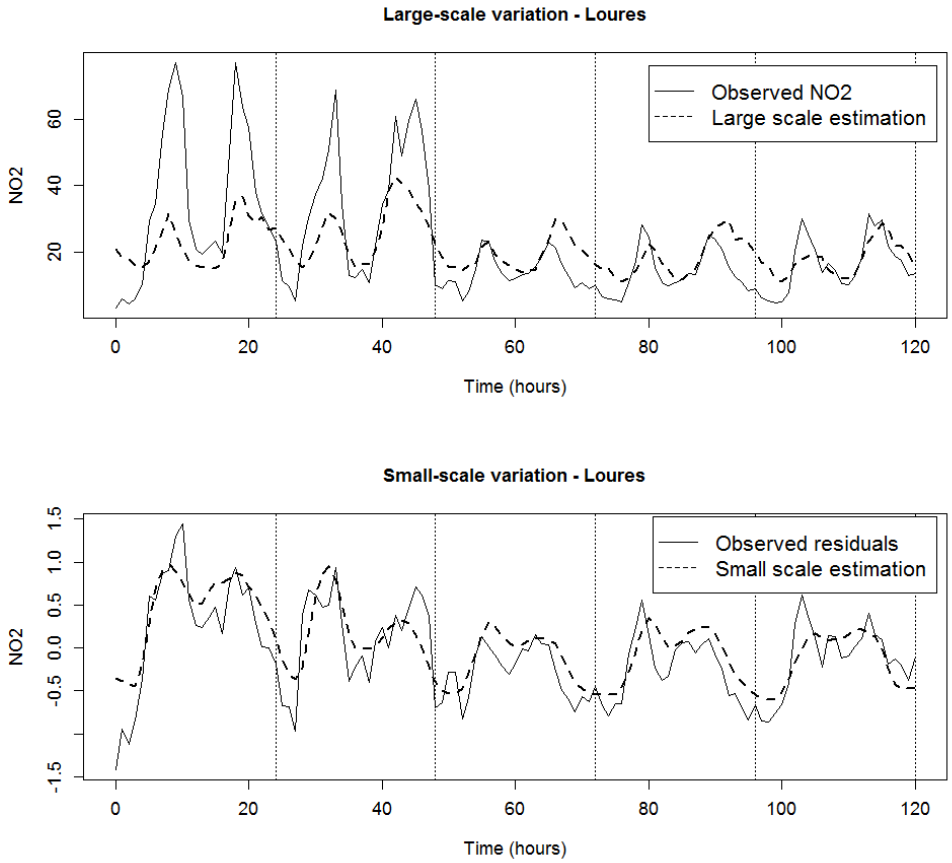


Figure 3.7: Estimation of the large-scale variation (top) and small-scale variation (bottom) of NO_2 concentrations in Loures Station from 2014-10-13 (Monday) to 2014-10-17 (Friday).

space-time point within the observation domain. Thus it allows to assess how pollution patterns change over space and time, as well as extending the current sampling design to locations without monitoring stations. This is illustrated in Figure 3.8 which represents the predicted spatio-temporal NO_2 concentrations process (small-scale variation) over Portugal on a Friday and a Sunday at 8:00, 13:00 and 18:00. We choose these days because Friday and Sunday are the days of the week with the highest and lowest concentration levels, respectively, while the choice of the times correspond to daily maxima, 8:00 and 18:00 and minimum, 13:00. Note that most of the temporal patterns in NO_2 concentrations result from anthropogenic activities and are captured by the mean or large-scale variation. The first comment is that NO_2 concentrations present a

strong spatial pattern that does not present much variation over time: along the day and over different days. The space-time residual process achieves higher values on the coast where most of the urban and traffic monitoring station are located, corresponding to higher population density.

One may find slight differences on spatial patterns in interior zones of Portugal probably justified by the lack of monitoring stations, becoming harder to produce accurate estimations. Moreover, we can conclude that the estimated residuals slightly decrease, when comparing Friday and Sunday, mainly at 8:00 and 18:00. This should be explained by the lower traffic typical from weekends at these moments of the day.

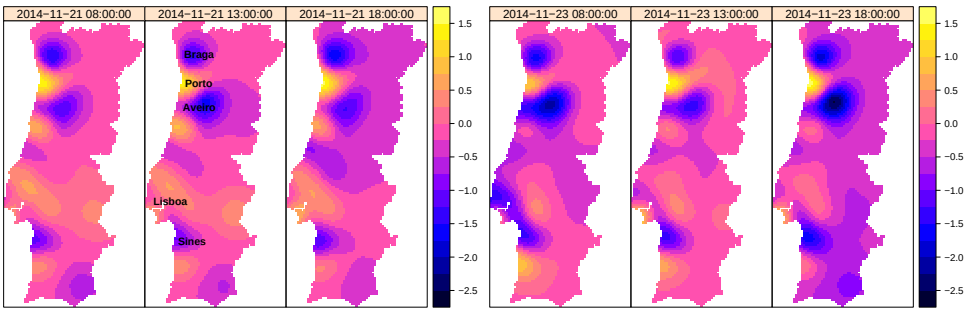


Figure 3.8: Space kriging maps for 2014-11-21 (Friday) and 2014-11-23 (Sunday), aiming to estimate the intra- and inter-day spatial patterns of NO_2 after removing the estimated trend.

A further application of space-time kriging allows to predict missing values in a specific station. These missing values may occur occasionally at some time points or when the station becomes inactive. Firstly, to illustrate this application, we proceed with the estimation of large and small-scale variation from Monday 2014-10-06 to Friday 2014-10-10 for Vila Nova da Telha, a suburban and background station from Maia county with no observations during this period. The results are presented at Figure 3.9, dashed lines, in the top panel, represent the 95% confidence bands for the estimated large-scale variation obtained by moving block bootstrap in time dimension, as explained in Section 3. The 95% confidence bands for the estimated small-scale variation, in the bottom panel, were obtained using kriging tools. We note that the estimated

afternoon peak seems to occur 1 hour later which might be explained by the fact that Maia is a satellite town of Porto, leading to a postponed rush hour traffic. Wednesday's NO_2 concentrations are lower with a somewhat different pattern from the remaining weekdays, which is also noted for other stations.

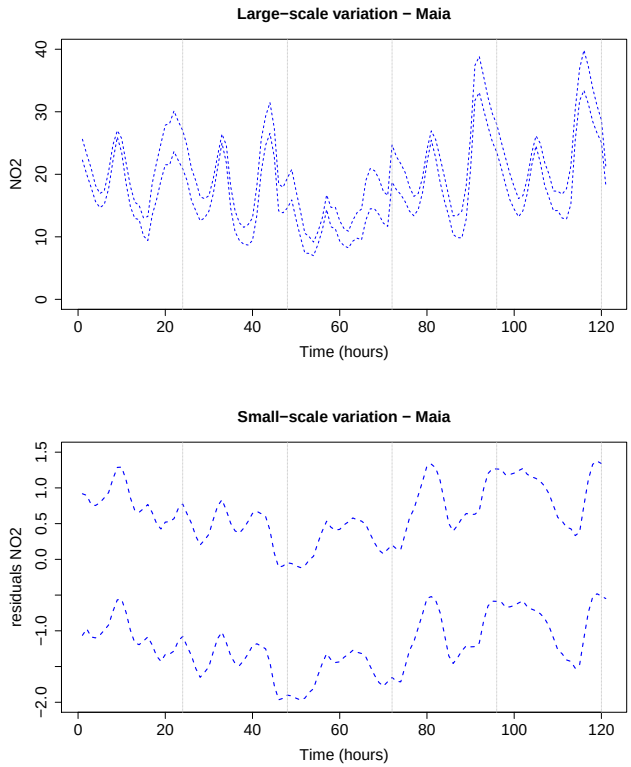


Figure 3.9: Estimation of the large-scale variation, top panel, and small-scale variation, bottom panel, of NO_2 concentrations in Maia station from Monday 2014-10-06 to Friday 2014-10-10. The dashed-lines identify the 95% confidence bands for: large-scale variation, obtained by a moving block bootstrap, each block with 5 weeks sliding 3 hours, generating 456 replicates (top panel); small-scale variation obtained by kriging tools (bottom panel).

3.5.2 Forecasting

The proposed model and associated modelling strategy enables to produce forecasts for NO_2 and quantify the associated uncertainty, as well as to analyse scenarios of possible future situations such as climate change and environmental policies. As explained

before, the NO_2 forecasts are acquired through the mean predictor.

In Portugal, December 2015 was considered atypically warm with a mean temperature of 11.8^0C, the second warmest since 1931. Consider the 14th of December, a Monday, with mean valued for temperature, wind speed and relative humidity 16.1^0C, 15.6 km/h and 87%, respectively. The daily mean forecasts for NO_2 in the 39 stations are represented in the right panel of Figure 3.10. The point estimates are classified for easiness of representation. Since QualAr NO_2 levels for December 2015 are not available at the time of writing, we compare these forecasts with fitted NO_2 levels for Monday 15th December 2014, left panel of Figure 3.10, a day with somewhat different meteorological conditions: mean temperature of 11.2^0C, wind speed of 9.9 km/h and relative humidity of 78.5%.

As expected, due to the altered weather conditions in 2015, the predictions of NO_2 levels for this year are lower than for 2014, in particular in the north of the country.

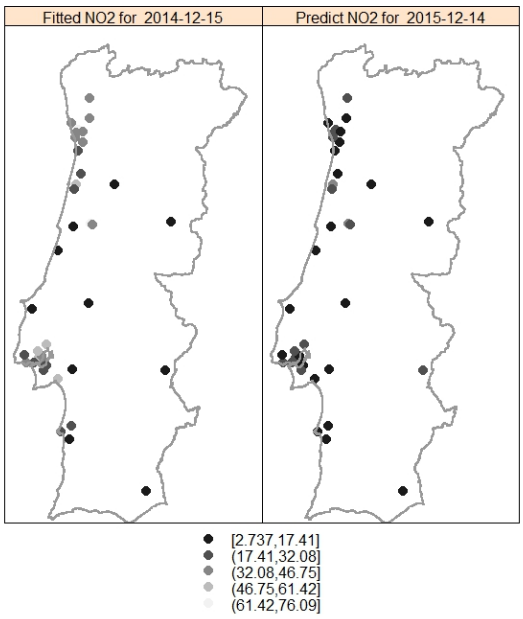


Figure 3.10: Daily mean of fitted NO_2 levels for 2014-12-15 (left). As meteorological data are available earlier than NO_2 levels, predictions for NO_2 levels for 2015-12-14 (right).

To further analyse the impact of meteorological variables (wind speed and relative humidity), we now compare hourly NO_2 concentrations observed during a week in 2014

with the corresponding 2015 forecasts for the same weekdays. The analysis is illustrated in Vila do Conde, a suburban and background station, between 15th and 21st December of 2014 (14th to 20th December of 2015). In Figure 3.11, all the meteorological variables and NO_2 levels for 2014 represent observed values, while the bottom right panel represents NO_2 forecasts for 2015. Bearing in mind that in 2014 the values of wind speed ranged from 0 to 15 km/h and in 2015 ranged from 0 to 30 km/h (top panels), and the increased variability of relative humidity in 2015 (middle panels), we note a significant decrease in the forecasts of NO_2 concentrations for 2015. Furthermore, the maximum peaks in the wind speed correspond to the minimum peaks of NO_2 concentrations, showing a "mirror" alike effect.

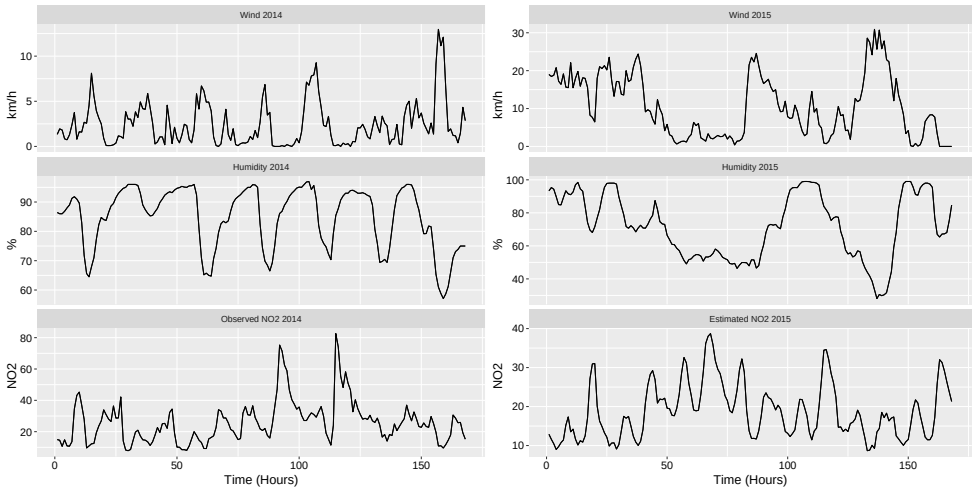


Figure 3.11: Observed NO_2 concentrations and meteorological variables in Vila do Conde, suburban and background station from 2014-12-15 (Monday) to 2014-12-21 (Sunday) (left). Meteorological variables in Vila do Conde from 2015-12-14 (Monday) to 2015-12-20 (Sunday) (right) and corresponding NO_2 forecasts.

3.5.3 Scenario analysis

Scenario analysis is achieved with conditional forecasting in which future (unknown) realizations of the explanatory variables are fixed at plausible values of interest. To illustrate the potential of the model in scenario generation, we obtain NO_2 forecasts under two distinct scenarios: if wind speed duplicates, and if relative humidity is re-

duced by half. In particular, we choose again Vila do Conde station, as being located in the north Portuguese coast, typically a windy and humid region. Figure 3.12 displays the observed NO_2 concentrations from 2014-12-12 (Monday) to 2014-12-18 (Sunday), against the NO_2 forecasts under the two scenarios which are being considered. The results confirm that an increase in wind speed provokes, in general, a decrease in NO_2 concentrations and a decrease in relative humidity leads, generally, an increase in NO_2 levels.

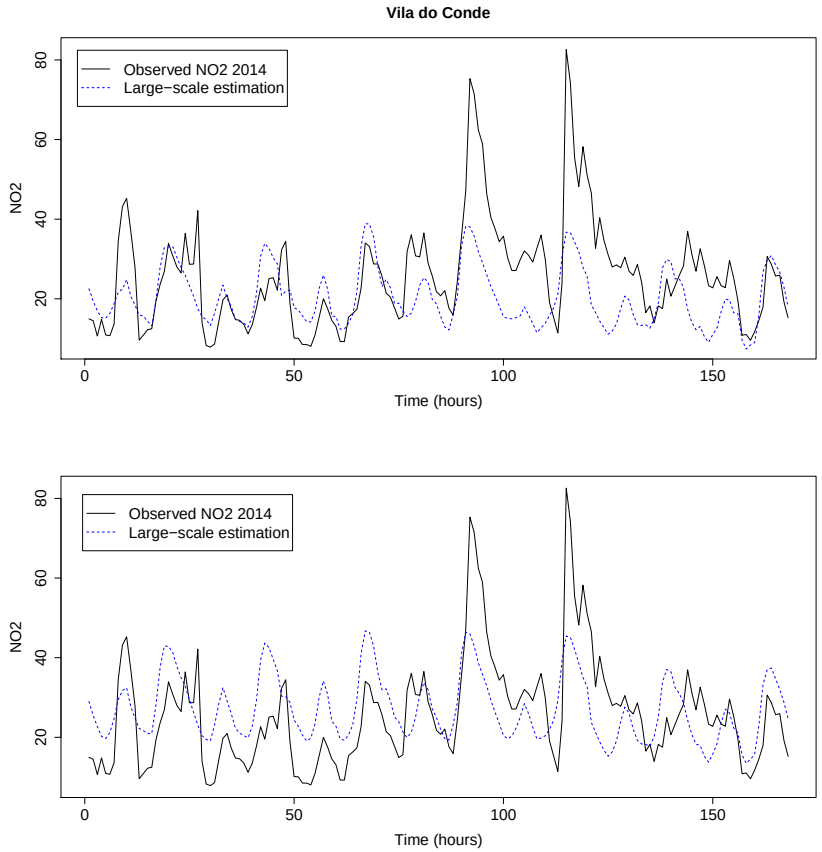


Figure 3.12: Observed NO_2 concentrations, in Vila do Conde station, from 2014-12-12 (Monday) to 2014-12-18 (Sunday). The dashed-lines represent NO_2 forecasts under the scenarios: wind speed duplicates (top panel) and relative humidity reduced by half (bottom panel).

A last example of scenario generation is the enforcement of environmental policies

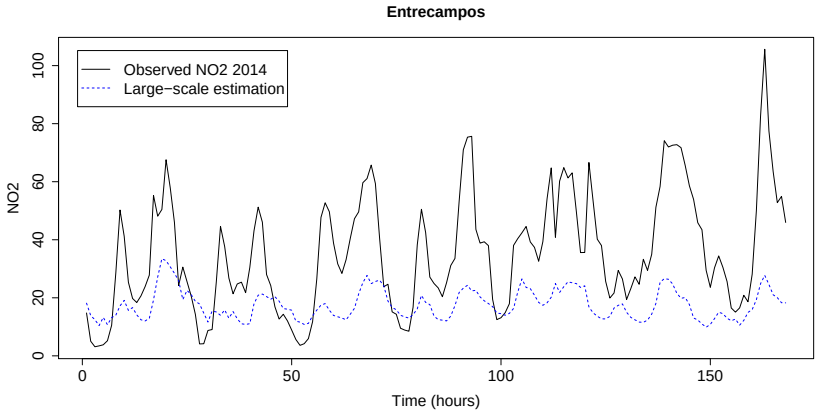


Figure 3.13: Observed NO_2 concentrations in Entrecampos station from 2014-12-12 (Monday) to 2014-12-18 (Sunday). The dashed-line represents NO_2 forecast under the scenario of changing this station from traffic to background classification.

that many European cities are taking by pondering the permanent prohibition of vehicles in certain areas. This is equivalent to changing the type of site of a station located in a city from traffic to background. To illustrate this situation we consider Entrecampos which is an urban and traffic station located in Lisbon, where only vehicles registered after 1996 can circulate. Figure 3.13 displays the observed NO_2 levels together with NO_2 forecast if Entrecampos station becomes classified as background, assuming that the meteorological variables are the same as in 2014. The decrease not only in mean but also in variability of NO_2 levels is noteworthy.

3.6 Conclusions

In this Chapter, an easily implementable two-step approach is suggested to model spatial and high resolution temporal data, which allows inference on the large-scale and small-scale variation components of the spatio-temporal stochastic process. The framework is particularly useful when data exhibits multiple seasonal patterns imposed by social habits, anthropogenic activity and natural cycles explained by meteorological condition, simultaneously incorporating any additional information considered relevant to explain the phenomenon. This work contributes to the characterization of the space-time dynamics, which can be used to complement the current sampling design by space-

time prediction, to obtain forecasts and perform scenario analysis in environmental data as NO_2 concentrations, as well as in other data sets with similar characteristics, such as electrical demand. The proposed modelling approach assumes data regular in time.

In the remaining Chapters of this thesis, we discuss model-based approaches to deal with data collected irregularly in time under Preferential Sampling schemes.

4

Modelling Preferential Sampling in time

Part of the work included in this Chapter is accepted for publication on REVSTAT - Statistical Journal, through a manuscript entitled "Modelling Irregularly Spaced Time Series under Preferential Sampling".

A particular case of irregularly spaced time series is that in which the sampling procedure over time depends also on the observed values. In such situations, there is stochastic dependence between the process being modeled and the times of the observations. In this Chapter, we introduce the concept of Preferential Sampling in the temporal dimension and we propose a model-based approach to make inference and prediction. We first consider a Monte Carlo approach for maximum likelihood estimation of the model and then we consider a numerical method based on a Laplace approach to optimize the likelihood.

4.1 Introduction

Analysis of experimental data that have been observed at different points in time leads to specific problems in statistical modelling and inference. In traditional time series the main emphasis is on the case when a continuous variable is measured at discrete equispaced time points, Tómasson (2015). There is an extensive body of literature on analysing equally spaced time series data, see for example Box *et al.* (2015) and Brockwell & Davis (2002). Nevertheless, unevenly spaced (also called unequally or

irregularly spaced) time series data naturally occur in many scientific domains. A particular case of irregularly spaced data is that in which the collection procedure along time depends also, for practical constraints, on the observed values. For example, a certain health indicator for an individual may be measured at different time points and with different frequencies depending on his health state. In a completely different setting, the times of occurrence of transactions in the financial markets depend largely on the value of the underlying asset. In environmental monitoring applications, or in the context of smart cities if it is decided to monitor more frequently when a value considered critical to human health is exceeded. Therefore, additional information on the phenomena under study is obtained from the frequency or time occurrence of the observations. Such situations in which there is stochastic dependence between the process being modeled and times of the observations may be coined as temporal Preferential Sampling, following Diggle *et al.* (2010) in the context of spatial statistics.

Preferential Sampling in time could be seen as a version of informative follow-up in longitudinal studies, see, for example, Lin *et al.* (2004), Ryu *et al.* (2007) and Liang *et al.* (2009) who proposed joint modelling and analysis of longitudinal data with possibly informative observation times via latent variables. In these studies the follow-up time process is considered dependent on the longitudinal outcome process and it should not be regarded deterministic in the design of the study. The analogous problem in the context of longitudinal clinical trial data has been studied too in the context of issues concerning missing values and dropouts, in the sense that a missing observation conveys partial information about the value that would have been observed. See, for example, Diggle & Kenward (1994), Hogan & Laird (1997) and Daniels & Hogan (2008).

In this work, we propose a model-based approach to analyse a time series observed under Preferential Sampling. The suggested framework considers the observed time points as the realization of a time point process stochastically dependent on an underlying latent process (e.g. an individual health indicator or the underlying asset). This latent process is assumed as Gaussian without loss of generality.

The developed work falls within the scope of irregularly spaced time series and it applies theory of point processes, presuming that the time of the observations have been produced by some form of stochastic mechanism. In the next Section, an introduction to temporal point processes, based on Daley & Vere-Jones (2003) is developed.

4.2 Basic concepts in point process theory

Point processes are stochastic processes that are used to model events that occur at random intervals relative to the time axis or the space axis. Thus, we may consider two main types of point processes: temporal point processes and spatial point processes.

A temporal point pattern is basically a list of times of events. For example, the time of an earthquake in seismology or the time of an extreme asset return in financial applications. An essential tool for dealing with this kind of data is a stochastic process modelling the point patterns: a temporal point process. The term point is used since we may think of an event as being an instant and thus we can represent it as a point on the time line. For the same reason the words point and event can be used interchangeably.

4.2.1 The Poisson point process

The Poisson point process is one of the most used and studied point process models due to its particularly convenient properties. In addition, it serves as a basis for the construction of more complicated models. There are two different types of Poisson processes: the homogeneous or stationary and the nonhomogeneous or inhomogeneous. The first one is the simplest model in point processes theory and is completely defined by

$$Pr\left\{N\left(a_i, b_i\right] = n_i, i = 1, \ldots, k\right\} = \prod_{i=1}^{k} \frac{\left[\lambda(b_i - a_i)\right]^{n_i}}{n_i!} e^{-\lambda(b_i - a_i)} \tag{4.1}$$

where $N\left(a_i, b_i\right]$ denotes the number of events of the process falling in the half open interval $(a_i, b_i]$ with $a_i < b_i \leq a_{i+1}$.

Equation (4.1) embodies three important properties:

1. the number of points in each finite interval $(a_i, b_i]$ follows a Poisson distribution with mean $\lambda(b_i - a_i)$;

2. the number of points in disjoint intervals are independent random variables;

3. the distributions are stationary: they depend only on the lengths $b_i - a_i$ of the intervals.

Thus, the joint distributions are multivariate Poisson of the special type in which the variates are independent.

The constant λ in (4.1) is the characteristic parameter, called the intensity or point density, of the homogeneous Poisson process.

The inhomogeneous Poisson point process is an extension of the simplest one, where the intensity λ is now a function of the time, $\lambda(t)$. The process can be defined exactly as in (4.1) with $\lambda(b_i - a_i) = \int_{a_i}^{b_i} \lambda dx$ replaced by

$$\Lambda(a_i, b_i] = \int_{a_i}^{b_i} \lambda(x)dx \tag{4.2}$$

The properties of the homogeneous Poisson process have natural analogues in the inhomogeneous case. Thus, the joint distributions are still Poisson, and the independence property still holds.

Sometimes the intensity function $\lambda(t)$ is rather irregular so it may be useful to consider instead the point density distribution function, Ghorbani *et al.* (2006). Suppose that there are n observations on $(0, T]$ at time points $t_1, ..., t_n$, in particular the conditional distributions are independently distributed on $(0, T]$ with a common distribution having density function $\frac{\lambda(t)}{\Lambda(0,T]}$. Consequently, the corresponding location density function of the n points is

$$f_n(t_1, \ldots, t_n) = \prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda} \tag{4.3}$$

where $\Lambda = \int_T \lambda(t)dt$ i.e. the n points form a sample of n independent points with a probability density function proportional to $\lambda(t)$.

4.2.2 Cox (doubly stochastic Poisson) processes

In our work we wish to model aggregated temporal point patterns where the aggregation is due to some stochastic heterogeneity induced by an unobserved process. This leads to a class of inhomogeneous Poisson processes with stochastic intensity functions, called the Cox processes. A Cox process can be regarded as the result of a two-stage random mechanism and for this reason Cox processes are sometimes termed "doubly stochastic Poisson process". In the first step, a non-negative intensity function $\lambda(t)$ is generated. Conditional on this, an inhomogeneous Poisson process with intensity function $\lambda(t)$ is constructed in the second step. In order words, given $\lambda(t)$, the point distribution

is completely random. This approach is a special case of the hierarchical modelling approach, which is commonly used in model construction in many areas of classical statistics.

In a stationary Cox process the intensity function is replaced by a stationary stochastic process with non-negative values. The realizations of this process are functions which are treated as intensity functions of inhomogeneous Poisson processes. All distributional properties of the point process generated are inherited from the stationary stochastic process, yielding a stationary point process model.

Formally, the Cox point process model is defined in two steps:

- Consider a stationary non-negative valued stochastic process $\{\Lambda(t) : t \in \mathbb{R}\}$.
- Given a realisation of the stochastic process, i.e. given that $\Lambda(t) = \lambda(t)$ for all $t \in \mathbb{R}$, the points of the corresponding realisation of the Cox process form an inhomogeneous Poisson process with intensity function $\lambda(t)$.

Assuming Gaussian data, we shall consider log-Gaussian Cox processes, i.e Cox processes where the logarithm of the intensity is a Gaussian process. Considering a Gaussian process $Z(t)$, this type of process cannot be used as the intensity of a Cox process since it can take negative values. Thus, a suitable transformation has to be applied to yield a Cox process. A very elegant transformation, resulting in a mathematically tractable model, is

$$\Lambda(t) = \exp(Z(t))$$

The corresponding process is termed a log-Gaussian Cox process and it was first described in Rathbun (1996) and Møller *et al.* (1998).

4.2.3 Simulation of a Poisson process

An useful simulation technique to simulate an inhomogeneous Poisson point process is *independent thinning*, that consists of using some predefined rules to remove points of a process and form a new one. Suppose that X is a Poisson process with intensity function $\lambda(t)$, and that each point of X is either deleted or retained, independently of other points. If the retention probability is $p(t)$, then the resulting process of retained points is Poisson with intensity $\lambda(t)p(t)$. Then, to simulate an inhomogeneous Poisson

point process with intensity $\lambda(t)$ is enough to start by simulating a homogeneous one with intensity $\lambda = \max_{t \in T} \lambda(t)$ and perform a thinning with retention probabilities (thinning function) $p(t) = \frac{\lambda(t)}{\lambda}$.

In practice, this means that, based on the thinning function, a decision is made for each point $t_1, \ldots, t_n$ in the sample of the homogeneous Poisson process with intensity λ as to whether to "retain" or "thin" it. A point t_i is retained with probability $p(t_i) = \frac{\lambda(t_i)}{\lambda}$, each point being retained or deleted independent of what happens to any of the other points.

The Cox processes can be simulated in a straightforward way, based on the hierarchical nature of the model. In a first step the intensity $\lambda(t)$ is generated and in a second step the point pattern is simulated given $\lambda(t)$ using the same method as for inhomogeneous Poisson processes.

4.3 A model for Preferential Sampling in time

We consider data obtained by irregularly sampling a continuous time phenomenon $S(t) : t > 0$ at a discrete set of times t_i, $i = 1, \ldots, n$. In many situations, $S(t)$ cannot be measured without error, hence, if $Y_i = Y(t_i)$ denotes the measured value at time t_i, a model for the data takes the form:

$$Y(t) = \mu + S(t) + N(0, \tau^2), \ t > 0 \tag{4.4}$$

Thus, this model has a set of components which are detailed as follows.

- $S(\cdot)$ is a stationary Gaussian process with $E[S(t)] = 0$. We consider $S(\cdot)$ as a continuous time autoregressive process of order 1, CAR(1), defined in Section 2.3 that satisfies the differential equation

  $$dS(t) + \alpha_0 S(t)dt = dW(t)$$

  where, α_0 is the autoregressive coefficient and $W(t)$ is a Wiener process with variance parameter σ_w^2.

- $Y = (Y_1, \ldots, Y_n)^t$ is multivariate Gaussian with mean μ_Y and covariance matrix Σ_Y

$$Y = (Y_1, \ldots, Y_n)^t \sim MVN\left(\mu_y \mathbf{1}, \Sigma_y\right)$$

with

$$\mu_y = \mu\mathbf{1} \quad \text{and} \quad \Sigma_y = \frac{\sigma_w^2}{2\alpha_0} R_y(\alpha_0) + \tau^2 I_n$$

where I_n is the $n \times n$ identity matrix, $\mathbf{1}$ denotes the n-element vector of ones and $R_y(\alpha_0)$ has elements $r_{ij} = \rho\left(|t_i - t_j|\,; \alpha_0\right)$ defined by equation (2.10) from Section 2.3 ($\rho(h) = \exp(-\alpha_0\,|h|)$). An equivalent formulation is that conditional on $S(\cdot)$, the $Y(t_i)$ are mutually independent, normally distributed with mean $\mu + S(t_i)$ and common variance τ^2.

- $T = (t_1, \ldots, t_n)$ denotes a stochastic process of observation times.

Under the above mentioned assumption that the sampling times are stochastic, the joint distribution of S, T and Ymust be specified. Considering the stochastic dependence between S and T, the model to deal with Preferential Sampling is defined through $[S, T, Y]$ written as:

$$[S][T\,|S][Y\,|S(T)\,] \tag{4.5}$$

where $[\cdot]$ means "the distribution of", $S = \{S(t) : t > 0\}$, $T = (t_1, \ldots, t_n)$ and $S(T)$ represents $\{S(t_1), \ldots, S(t_n)\}$.

In this Chapter, we define a specific class of models through the additional assumptions

- Conditional on S, T is an inhomogeneous Poisson process with intensity

$$\lambda\,(t) = \exp\left\{a + \beta S\,(t)\right\} \tag{4.6}$$

- Unconditionally T is a log-Gaussian Cox process. The log-Gaussian Cox process, see Section 4.2.2, is a flexible class of point pattern models that allows conditioning the sampling times to the variable of interest. β is the parameter that controls the degree of preferentiality, for example, when $\beta > 0$ the sample times are concentrated, predominantly, near the maximum of the observed values and when $\beta = 0$ it corresponds to the situation of an homogeneous, non-preferential, sampling.

- Conditional on S and T, Y is a set of mutually independent Gaussian variates with τ^2 being the measurement error variance.

To obtain the parameters of the model we use maximum likelihood estimation. For the shared latent process model, the likelihood function for data T and Y can be expressed as

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [T, Y, S] dS = \int_S [S][T, Y|S] dS = \int_S [S][T|S][Y|T, S] dS \tag{4.7}$$

where $\boldsymbol{\theta} = (\mu, \sigma_w, \alpha_0, \tau, \beta)$ represents the set comprising the model parameters.

Although the construction of this model is driven by a Preferential Sampling context, it may be applied to model any type of irregularly spaced time series. One of its advantages is to make predictions at unobserved time points.

Prediction

The predicted value of $S(\cdot)$ at an unsampled time point $t_{n_i} < t_0 < t_{n_j}$, $S(t_0|T)$, is given by $S(t_0|T) = \mathrm{E}\big[S(t_0)|Y(T)\big]$. Considering that the process CAR(1) is Markovian, (Brockwell & Davis, 2002, p.358) shows that the conditional mean of $S(t_0)$ given $Y(T)$ is

$$\begin{aligned} S(t_0|T) &= \mathrm{E}\big[S(t_0)|Y(T)\big] \\ &= \exp\left(-\alpha_0(t_0 - t_{n_i})\right) Y(T) + \mu\left(1 - \exp\left(-\alpha_0(t_0 - t_{n_i})\right)\right) \end{aligned} \tag{4.8}$$

The variance of the prediction is

$$\sigma^2(t_0) = Var\big[S(t_0)|Y(T)\big] = \frac{\sigma_w^2}{2\alpha_0}\left(1 - \exp\left(-2\alpha_0(t_0 - t_{n_i})\right)\right) \tag{4.9}$$

4.4 Inference - Monte Carlo approach

4.4.1 Maximum likelihood estimation

Evaluation of the conditional distribution $[T|S]$ in (4.7) strictly requires the realization of S to be available at all $t \in T$. We consider a discretization of the S process with N

points and replace the exact locations T by their closest points on the grid. We then partition S into $S = \{S_0, S_1\}$, where S_0 denotes the values of S at each of n times $t_i \in T$, and S_1 are the values of S at the remaining $(N - n)$.

An algebraic simplification of $[Y|T, S]$ is $[Y|S_0]$ so, we can rewrite the integral in (4.7) as

$$L(\boldsymbol{\theta}) = \int_S [S][T|S][Y|S_0]\frac{[S|Y]}{[S|Y]}dS \tag{4.10}$$

Considering that $[S] = [S_1, S_0] = [S_1|S_0][S_0]$ and replacing the term $[S|Y]$ in the denominator of expression (4.10) by $[S|Y] = [S_0, S_1|Y] = [S_1|S_0, Y][S_0|Y] = [S_1|S_0][S_0|Y]$, equation (4.10) becomes

$$\begin{aligned} L(\theta) &= \int_S [S_1|S_0][S_0][T|S][Y|S_0]\frac{[S|Y]}{[S_1|S_0][S_0|Y]}dS \\ &= \int_S [T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]dS \\ &= E_{S|Y}\left[[T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right] \end{aligned} \tag{4.11}$$

Taking into account that the conditional expectation in (4.11) can be approximated by Monte Carlo, Maximum Likelihood Estimates (MLE's) are obtained by maximizing the Monte Carlo likelihood

$$L_{MC}(\boldsymbol{\theta}) = m^{-1}\sum_{j=1}^{m}[T|S_j]\frac{[Y|S_{0j}]}{[S_{0j}|Y]}[S_{0j}] \tag{4.12}$$

where S_j are assumed as realizations of the distribution of S conditional on Y. S_{0j} denotes the values of S_j restricted to the n observed time points. We may notice that j takes a value from 1 to m, the total number of Monte Carlo replicates. With this purpose, we use a technique known as conditioning by kriging, Rue & Held (2005), and the following construction. The new sample

$$S_j = U + \Sigma_S A^T \left(A\Sigma_S A^T + \tau^2 I_n\right)^{-1}(V - AU)$$

where A is the $n \times N$ matrix whose ith row consists of $N-1$ 0s and a single 1 to identify the position of t_i within $T = (t_1, \ldots, t_n)$; $U = \Sigma_S^{1/2} u \sim MVN(0, \Sigma_S)$ with $u \sim N(0,1)$ and $\Sigma_S^{1/2}$ is obtained from the Cholesky decomposition and $V \sim MVN(y, \Sigma_Y)$.

Then S_j has the required multivariate Gaussian distribution of S given $Y = y$. In practice, to reduce Monte Carlo variance, we use antithetic pairs of realizations, i.e. for each $j = 1, \ldots, m/2$ set $S_{2j} = 2\mu_c - S_{2j-1}$, where μ_c denotes the conditional mean of S given Y, Diggle *et al.* (2010).

$T|S_j$ in (4.12) is an inhomogeneous Poisson process with intensity given by equation (4.6)

$$\lambda(t) = \exp(a + \beta S_j(t))$$

As we have seen in Section 4.2.1, the density function is given by $\prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda}$. Consequently, and working with logarithm for computational reasons,

$$\log([T|S_j]) = \sum_{i=1}^{n} (a + \beta S_j(t_i)) - n \log\left(\int_0^T \exp(a + \beta S_j(t)) dt \right) \tag{4.13}$$

Since the S_j replicate is not known in $[0, T]$ domain, we can not calculate the integral, so we approximate the integral using the composed trapezium formula for unequally spaced data.

$$\int_0^T \exp(a + \beta S_j(t)) dt = \frac{1}{2} \sum_{k=1}^{N-1} (t_{k+1} - t_k)(\exp(a + \beta S_j(t_{k+1})) + \exp(a + \beta S_j(t_k)))$$

$[S_{0j}]$ in (4.12) is multivariate Gaussian with mean 0 and covariance matrix $\Sigma_{S_{0j}}$

$$S_{0j} \sim MVN\left(0, \Sigma_{S_{0j}}\right)$$

with

$$\Sigma_{S_{0j}} = \frac{\sigma_w^2}{2\alpha_0} R_{S_{0j}}(\alpha_0)$$

where $R_{S_{0j}}(\alpha_0)$ is the $n \times n$ correlation matrix with elements $r_{ik} = \rho\left(|t_i - t_k|\,; \alpha_0\right)$ defined by equation (2.10) from Section 2.3.

$[S_{0j}|Y]$ in (4.12) is multivariate Gaussian with mean $\mu_{S_{0j}|Y}$ and covariance matrix $\Sigma_{S_{0j}|Y}$, so that

$$\mu_{S_{0j}|Y} = \Sigma_{S_{0j}}\Sigma_Y^{-1}(y - \mu\mathbf{1})$$

$$\Sigma_{S_{0j}|Y} = \Sigma_{S_{0j}} - \Sigma_{S_{0j}}\Sigma_Y^{-1}\Sigma_{S_{0j}}^t$$

For more details about conditional distribution see for e.g. Anderson (1984).

In equation (4.12), $[Y|S_{0j}] = \prod_{i=1}^n [Y_i|S_{0j}(t_i)]$ with $Y_i|S_{0j}(t_i) \sim N(\mu + S_{0j}(t_i), \tau^2)$, meaning that, conditional on S and T, Y is a set of mutually independent Gaussian variates.

Obtained the MLE's we can plug them into (4.8) and (4.9), treating them as known. We are in position of doing the so-called plug-in predictions.

4.4.2 Numerical studies

In this Section, we document the performance of the model with time series simulated under preferential and non preferential (irregular and regular sampling) scenarios. The simulation allows to control the degree of preferentiality. We compare the results from our model with the traditional Kalman filter approach to irregularly spaced data (cts package (Wang, 2013)). To simulate a time series under Preferential Sampling, we use the procedure described in Section 4.2.3.

Simulation design

To generate a time series under Preferential Sampling, we first generate a realization of S from model (4.4) with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$, discretized in 400 equally spaced time points. These values correspond to $Var[S(\cdot)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ and $\phi = \frac{1}{\alpha_0} = 5$, being the latter related to the lag beyond which there is no correlation for practical purposes. To generate Y from model (4.4), we consider $\mu = 0$ and $\tau = 0.1$, conducting three separate sampling procedures over the realization of S

- Preferential Sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (4.6) and $\beta = 2$;

- irregular sampling: we obtain $n = 70$ sampling times T from (4.6) and with $\beta = 0$, illustrating the situation without Preferential Sampling;
- regular sampling: we obtain $n = 70$ sampling times with equidistant observations.

To illustrate the results of these sampling schemes, we represent in Figure 4.1 a realization of the process S (gray line) and the three resulting data sets. We have 70 sampling times (black points), considering $\beta = 2$ in the process intensity function, in which the preferential nature of the sampling process results in sample times falling predominantly near the maxima. For 70 sampling times (white points), we consider $\beta = 0$, the situation without Preferential Sampling and with irregularly sampling points. For the remaining 70 points (star points), we have the situation of regular spaced sampling times.

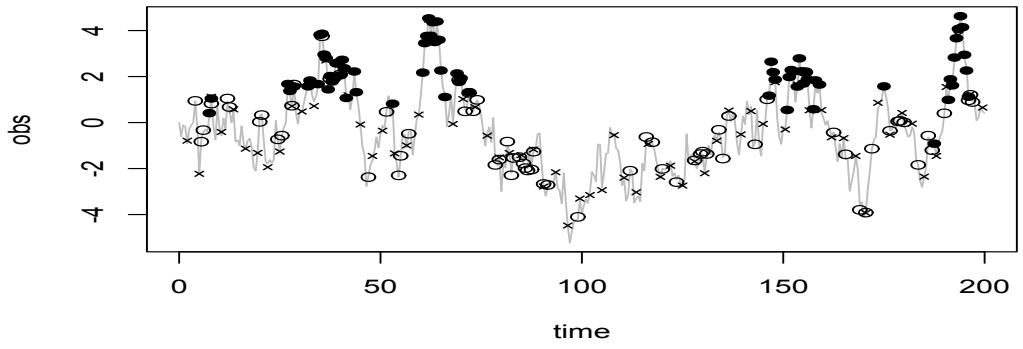


Figure 4.1: Sample times with Preferential Sampling nature (black points), without Preferential Sampling and irregularly spaced time points (white points), regular spaced time points (star points) and underlying process S (gray line).

Estimation results

The parameters μ, σ, ϕ, τ and β are the target of estimation. The estimates are obtained under (4.12), henceforward denoted by MCMLE's and from the Kalman filter, denoted by MLE's. For the maximization of our Monte Carlo log-likelihood function, we considered a total of grid points $N = 400$ and a total number of replicates $m = 1000$.

The results of the mean and standard errors of each parameter, obtained from a total of 250 independent samples are summarized in Table 4.1.

		PS Data set ($\beta = 2$)		Irregularly Sampling ($\beta = 0$)		Regular sampling	
	True	PS model	CTS	PS Model	CTS	PS Model	CTS
$\widehat{\mu}$	**0**	**0.13** (0.18)	**1.91** (0.51)	**0.04** (0.12)	**-0.01** (0.38)	**0.02** (0.22)	**-0.01** (0.37)
$\widehat{\sigma}$	**1.58**	**1.53** (0.21)	**0.91** (0.16)	**1.64** (0.11)	**1.54** (0.21)	**1.60** (0.13)	**1.47** (0.24)
$\widehat{\phi}$	**5**	**5.71** (1.01)	**2.52** (1.89)	**5.20** (0.48)	**5.41** (2.03)	**5.12** (0.89)	**6.56** (3.42)
$\widehat{\tau}$	**0.1**	**0.12** (0.04)	**0.17** (0.10)	**0.11** (0.01)	**0.22** (0.14)	**0.11** (0.02)	**0.56** (0.32)
$\widehat{\beta}$	**2 or 0**	**1.76** (0.39)		**0.00** (0.07)		**0.00** (0.02)	

Table 4.1: Maximum likelihood estimates, under PS model (MCMLE's) and by cts package (MLE's), mean (standard errors) obtained from a total of 250 independent samples.

By analysing Table 4.1, we conclude that the model for Temporal Preferential Sampling, when the sample times are preferentially sampled, presents estimates for the parameters less biased and shows considerable success, particularly in estimating mean (μ) and β parameters. When the preferentiality degree is null, with regular and irregularly sampling the estimation methods are essentially equivalent.

Further studies with β taking non-integer and negative values (sampling times are concentrated, predominantly, near the minimum of the observed values) lead to similar conclusions.

Inference on $S(t)$

To illustrate the potential of the model-based approach, we obtain prediction confidence intervals for the underlying process $S(t)$. For this purpose, MCMLE's and the MLE's from Kalman filter are plugged-in equation (4.8) to predict $S(t)$ at equally spaced time points. These together with the corresponding standard errors, in (4.9), allowed us to calculate prediction 95% confidence intervals and estimate their coverage.

Figure 4.2 represents one simulation of $S(t)$ (black line), the corresponding Preferential Sampling data (black points) and the predictions at equally spaced time points acquired from MCMLE's (white points) and MLE's (gray points). MLE's which do not take into account the preferential character of the data lead to predictions with larger bias (overestimation of the observations) and smaller variance than that of MCMLE's.

To analyse the impact of ignoring Preferential Sampling on the quality of predictions, we conducted a second simulation study. We simulated 250 realizations of S, for

each one we constructed a Preferential Sampling data set and predict $S(t)$ at 50 equally spaced time points. In fact, in the overall simulation results, confidence intervals from MCMLE's present an estimated coverage of 88% while the MLE's provide an estimated coverage of just 73%. Thus, the proposed model leads to estimates that are less biased but with larger variance, reflecting the uncertainty associated with the observations.

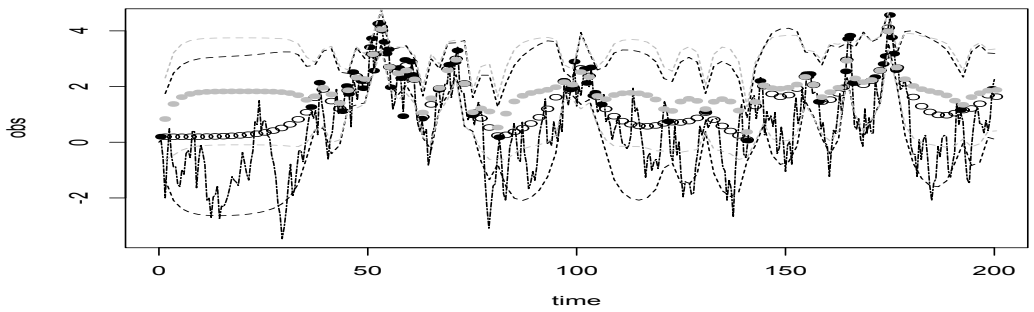


Figure 4.2: Predictions acquired from MCMLE's (white points) and MLE's (gray points), dashed line are confidence bands, black points are the Preferential Sampling data and black line is the underlying process S .

Numerical studies suggest that our model is effective at detecting potential Preferential Sampling situations, estimating an adequate model and obtaining predictions for the process.

4.4.3 Application to real data

In this Section, we apply our modelling procedure to two real data sets. The first consists of a time series related to the biomedical marker level of platelet after a cancer patient undergoes a bone marrow transplant, while the second is a data set associated with measurements of the lung function of an asthma patient.

Biomedical marker

We consider the problem of monitoring the level of a biomedical marker, platelet, after a cancer patient undergoes a bone marrow transplant. The data, 91 measurements in different days of log(platelet) [PLT] is represented in Figure 4.3 and is studied by Shumway & Stoffer (2017) as missing data problem. The first 35 data points correspond

to daily data. Afterwards the indicator began to show better results and observations become irregularly spaced. According to Jones (1984), "Platelet count at about 100 days post transplant has previously been shown to be a good indicator of subsequent long term survival". This data is available in the package astsa Stoffer (2017) with the name of "blood".

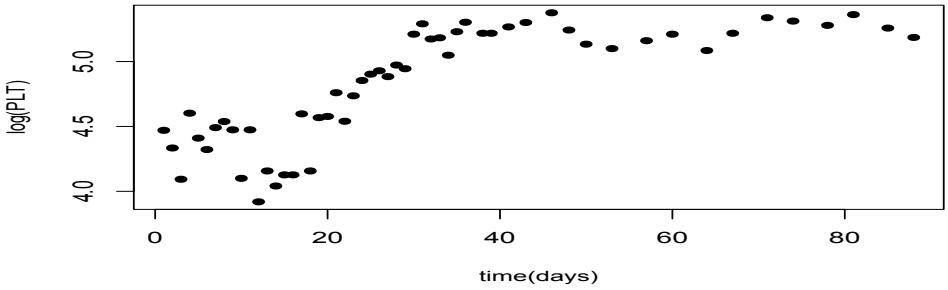


Figure 4.3: Measurements of the log(platelet) [PLT] .

The MCMLE's for model parameters are: $\hat{\mu} = 4.97$, $\hat{\phi} = 54.85$, $\hat{\sigma} = 0.52$, $\hat{\tau} = 0.14$ and $\hat{\beta} = -1.51$. MLE's from Kalman filter, are $\hat{\mu} = 4.89$, $\hat{\phi} = 53.94$, $\hat{\sigma} = 0.42$ and $\hat{\tau} = 0.13$. The estimated value for β with negative sign indicates that the data was, in fact, observed under a preferential framework whereby the patient is observed more frequently when the biomarker shows lower values. As expected, since the estimated value for β is negative, the mean (μ) estimated from MCMLE's is greater than the estimated from MLE's.

Predictions of the biomarker within the period of observations are obtained plugging-in the estimated parameters in equations (4.8) and (4.9). Figure 4.4 top panel shows the 95% prediction intervals for (log of) the biomarker obtainded from MCMLE's, while the bottom panel represents the 95% prediction intervals obtained from MLE's. As we saw in the numerical studies with simulated data, the predictions obtained from MCMLE's present larger variance reflecting the uncertainty associated with the preferential data under analysis.

This type of study allows greater knowledge of the underlying process and analyse, for example, when measurements of the patient's health indicator should have been carried out.

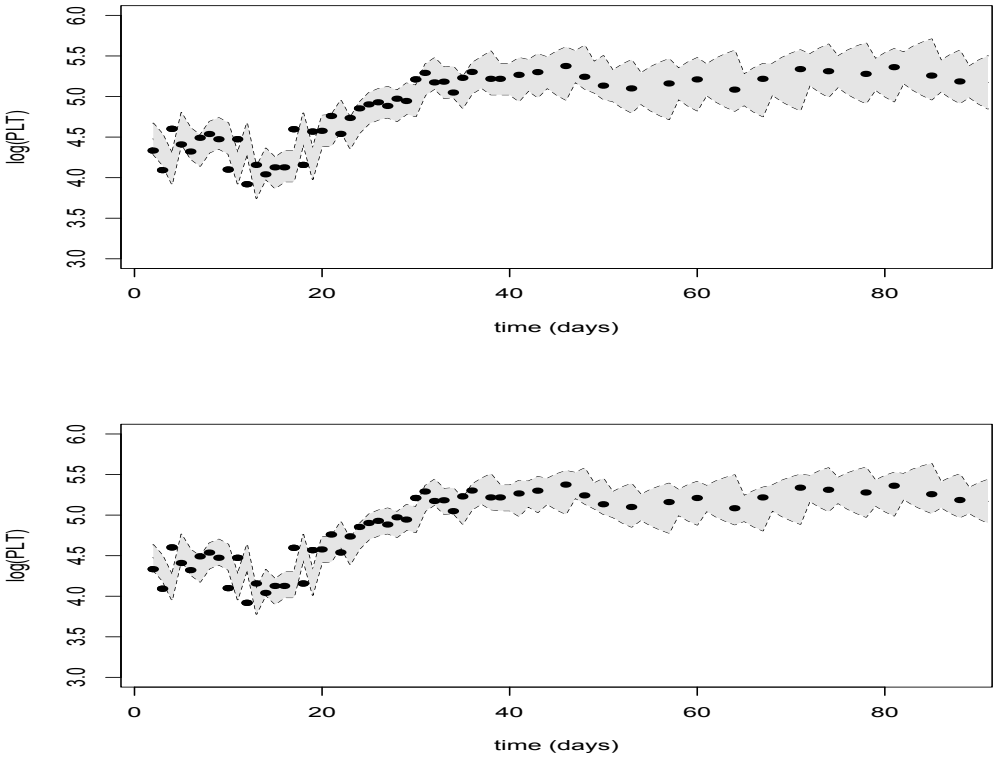


Figure 4.4: Prediction 95% confidence intervals using predictions acquired from MCMLE's (top) and MLE's (bottom).

Lung function of an asthma patient

Belcher *et al.* (1994) analysed 209 measurements of the lung function of an asthma patient. The time series is measured mostly at 2 hour time intervals but with irregular gaps as demonstrated by the unequal space of tick marks in Figure 4.5. This data is available in the package cts Wang (2013) with the name of "asth".

To assess the performance and the utility of the proposed model, we select the last 50 observations of "asth" data, corresponding to the period with more missing observations. We considered a log-transformation of the data which leads to more symmetric distribution of the values. We obtain predictions within the period of these observations, aiming to "complete" the data set. Figure 4.6 shows predictions of (log of) the variable of interest for that patient at regular time points. The MCMLE's for

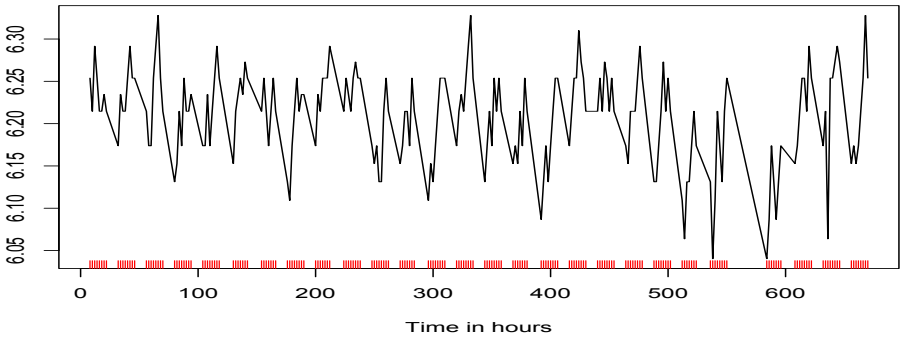


Figure 4.5: Measurements of the log (lung function).

model parameters are $\hat{\mu} = 6.18$, $\hat{\phi} = 2.83$, $\hat{\sigma} = 0.06$, $\hat{\tau} = 0.03$ and $\hat{\beta} = 0.62$. The small positive value for $\hat{\beta}$ reveals a week degree of preferentiality.

Additionally, since this data set covers about 30 days of a health state monitoring of the patient, we perform three further analyses. We obtain sub samples, with 50 time points, of the entire data set using the thinning algorithm described in Section 4.2.3, assuming β equals to -2, 0 and 2 in (4.6), illustrating the cases

- State of poor health ($\beta = -2$);
- State of healthy ($\beta = 2$);
- Random sample ($\beta = 0$)

We obtain the prediction confidence intervals for the underlying process, as described in Section 4.4.1, and the results for confidence intervals obtained from MCMLE's present an estimated coverage of 92% in both preferential samples ($\beta = \pm 2$) and 97% in the case of the random sample. These results help to justify the small degree of preferentiality present in data.

4.5 Inference - Laplace approach

In the first part of this Chapter, we have derived an MCMLE's algorithm for estimating the model parameters. The algorithm works well and allows the model to be used in

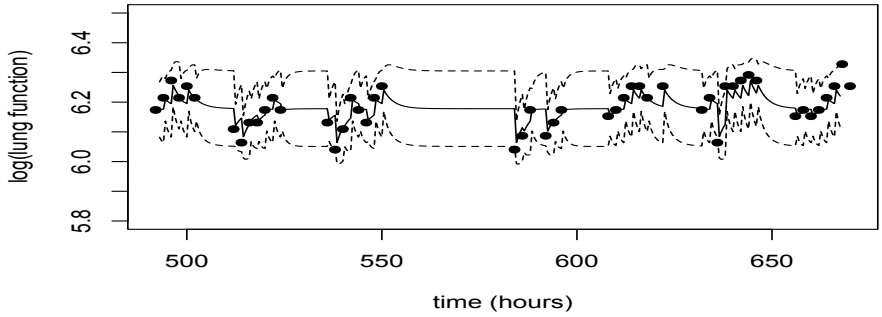


Figure 4.6: Predictions of (log of) the variable of interest (black line) and Confidence Intervals (dashed line). Black points are observations for the logarithm of lung function of an asthma patient.

practice, however, the convergence of this algorithm is very slow and the running time becomes burdensome for longer time series and a large number of Monte Carlo samples. Also note that the algorithm is sensitive to starting values $\theta_{\mathbf{0}}$. Besides these, the large variability between likelihood values in each Monte Carlo iteration makes the likelihood difficult to optimize. Our aim is now to

- Work directly with the likelihood function (4.7), using an alternative numerical method that uses the Laplace approximation for the marginal likelihood avoiding previous Monte Carlo approximation;
- Adopt a technique based on stochastic partial differential equation (SPDE) to approximate the CAR process. This allows to create a temporal mesh and corresponding components of the sparse precision matrix of a Gaussian Markov Random Field (GMRF) in time-dimension;
- Improve significantly the optimization of the likelihood function using programming language C++.

The above mentioned numerical techniques based on the Laplace approximation and SPDE have become usual when dealing with complex models and large data sets, Dinsdale & Salibian-Barrera (2018) and Diggle & Giorgi (2017). These changes will

hopefully result in a large increase in the stability of our parameter estimates, particularly in comparison with our previous method based on Monte Carlo approximation.

4.5.1 Methodological details

An alternative method (henceforth LAP) to the Monte Carlo simulation outlined in Section 4.4.1 is to utilize Automatic Differentiation of a Laplace Approximation to the marginal likelihood, to evaluate directly equation (4.7), i.e.

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [T, Y, S] dS = \int_S [S][T, Y|S] dS = \int_S [S][T|S][Y|T, S] dS$$

Automatic Differentiation

Automatic Differentiation (Griewank & Walther, 2008), also known as Computational Differentiation or Algorithmic Differentiation, is a set of techniques that numerically differentiates a function, which frees us from calculating and incorporating the derivatives. Two methods, "source transformation" and "operator overloading" are commonly used to implement automatic differentiation. CppAD (Bell, 2012), a package for C++ algorithmic differentiation, implements the "operator overloading" approach which is easier to implement and use compared with "source transformation". The R package TMB, short for Template Model Builder, (Kristensen *et al.*, 2016) uses CppAD to provide up to third order derivatives of the joint log-likelihood function. These derivatives are the required for the Laplace Approximation of the marginal likelihood.

Laplace Approximation

The Laplace approximation is used to approximate the integral in the likelihood (4.7). If we assume that the likelihood function for $L(\boldsymbol{\theta})$ can be written as

$$L(\boldsymbol{\theta}) = \int_S exp(-f(S, \boldsymbol{\theta})) dS \tag{4.14}$$

where $f(S, \boldsymbol{\theta})$ denote the negative joint log-likelihood of the data, $\boldsymbol{\theta}$ is the vector of parameters (fixed effects) and S the random effects. The Laplace approximation for $L(\boldsymbol{\theta})$ is

$$L^*(\boldsymbol{\theta}) = (2\pi)^{N/2} det(H(\boldsymbol{\theta}))^{-1/2} exp(-f(\widehat{S}(\boldsymbol{\theta}), \boldsymbol{\theta}))$$

where

$$\widehat{S}(\boldsymbol{\theta}) = arg_S min f(S, \boldsymbol{\theta}) \tag{4.15}$$

and $H(\boldsymbol{\theta})$ is the Hessian of f with respect to S evaluated at $\widehat{S}(\boldsymbol{\theta})$,

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial S^2} f(S, \boldsymbol{\theta})|_{S=\widehat{S}(\boldsymbol{\theta})}$$

The estimate of $\boldsymbol{\theta}$ minimizes the negative of the logarithm of the Laplace approximation,

$$-logL^*(\boldsymbol{\theta}) = -\frac{N}{2} log(2\pi) + \frac{1}{2} logdet(H(\boldsymbol{\theta})) + f(\widehat{S}(\boldsymbol{\theta}), \boldsymbol{\theta}) \tag{4.16}$$

This objective function and its derivatives acquired by using Automatic Differentiation, are required to apply standard nonlinear optimization algorithms (e.g., nlmimb) to optimize the objective function and obtain the estimate for $\boldsymbol{\theta}$. Using TMB library, the user has to define the joint log-likelihood of the data and (i.e. conditional on) the random effects as a C++ template function. The other operations such as integration and calculation of the marginal score function, are done directly in R language. The package evaluates and maximizes the Laplace approximation of the marginal likelihood where the random effects are automatically integrated out. This approximation, and its derivatives, are obtained using automatic differentiation (up to order three) of the joint likelihood. In the case of Preferential Sampling, we simply have to define the joint negative log-likelihood

$$f(S, \boldsymbol{\theta}) = -log([S][T|S][Y|S, T])$$

and allow TMB to integrate out the latent field S to evaluate approximately (4.7).

Uncertainty of the estimate $\widehat{\boldsymbol{\theta}}$ or of any differentiable function of the estimate $\zeta(\widehat{\boldsymbol{\theta}})$ that the user specifies, is obtained by the δ-method:

$$Var(\zeta(\widehat{\boldsymbol{\theta}})) = -\left\{ \frac{\partial \zeta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \left[\frac{\partial^2 (logL^*(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} \frac{\partial \zeta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{4.17}$$

These uncertainty calculations also require derivatives of (4.16). However, derivatives are not straight-forward to obtain using automatic differentiation in this context.

Stochastic Partial Differential Equation

To increase computational efficiency, we use a technique based on the stochastic partial differential equations (SPDE), Lindgren *et al.* (2011), to approximate the Gaussian process S. Lindgren *et al.* (2011) show that using an approximate stochastic weak solution to (linear) SPDE, for some Gaussian fields (GF) in the Matérn class, is possible to provide an explicit link between GF and Gaussian Markov randon fields (GMRF). Besides that we use the representation of a Gaussian process with Matérn covariance structure as the solution of the following SPDE,

$$\left(\phi^{-2} - \Delta\right)^{\alpha/2} (\omega S(t)) = W(t),\ t \in \mathbb{R}^+, \tag{4.18}$$

where $W(t)$ is Gaussian white noise, Δ is the Laplacian and ϕ is the range parameter of the Matérn covariance function $\gamma(u)$ in its standard parametrization,

$$\gamma(u) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} (u/\phi)^{\nu} K_{\nu} (u/\phi) : u \geq 0$$

where K_ν is the modified Bessel function of second kind and order $\nu > 0$ and σ^2 is the marginal variance. The integer value of ν determines the mean square differentiability of the underlying process, which matters for predictions made using such a model. However, ν is usually fixed since it is poorly identified in typically applications. The remaining parameters in (4.18) are $\alpha = \nu + 1/2$, from this we can identify the exponential covariance with $\nu = 1/2$, and ω that controls the variance,

$$\omega^2 = \frac{\Gamma(1/2)}{\Gamma(1)(4\pi)^{1/2}\phi^{-1}\sigma^2}$$

We approximate the process S by $\widetilde{S}$, where

$$\widetilde{S}(t) = \sum_{k=1}^{n} \psi_k(t) W_k,\ t \in \mathbb{R}^+$$

where $\psi_k(\cdot)$ are piecewise linear basis functions at a set of time knots and $W = W_1, ...W_n$ is a zero-mean multivariate Gaussian variate with covariance matrix Q^{-1}. The construction is done by projecting the SPDE onto the basis representation in what is essentially a Finite Element method. For $\alpha = 1$ the required form of Q is

$$Q = \omega^2(\phi^{-2}C + G_2)$$

where C and G_2 are sparse matrices whose explicit expressions can be found in Lindgren *et al.* (2011).

4.5.2 Numerical studies

A simulation study is performed to document the performance of LAP method. To simulate a time series under Preferential Sampling we use the procedure described in Section 4.2.3.

We first generate a realization of S from model (4.4) with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$, discretized in 800 equally spaced time points. These values correspond to $Var[S(\cdot)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ and $\phi = \frac{1}{\alpha_0} = 5$, being the latter related to the lag beyond which there is no correlation for practical purposes. To generate Y from model (4.4), we consider $\mu = 0$ and $\tau = 0.1$, conducting three separate sampling procedures over the realization of S

- Preferential Sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (4.6) and $\beta = 2$, which corresponds to the situation when the sampling times are concentrated, predominantly, near the maximum of the observed values;

- irregular sampling: we obtain $n = 70$ sampling times T from (4.6) and with $\beta = 0$, illustrating the situation without Preferential Sampling;

- Preferential Sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (4.6) and $\beta = -2$, which corresponds to the situation when the sampling times are concentrated, predominantly, near the minima of the observed values;

The parameters μ, σ, ϕ, τ and β are the target of estimation. We compare the parameter estimates, obtained from a total of 500 independent samples, for three alternative methods,

- LAP, implemented through C++, via package TMB;
- LAP, implemented via package INLA (Rue *et al.*, 2009);
- Kalman filter approach, implemented via package cts.

The INLA algorithm, proposed by Rue *et al.* (2009) and available in the R-INLA software package, is a numerical approximation method for Bayesian inference. INLA relies on Laplace approximation methods to numerically approximate posterior distributions. This method performs Gaussian approximations of the parameters by inferring their mode. Although posterior distributions do not necessarily have to be Gaussian, INLA relies on the fact that for most real problems and data sets, the conditional posterior of the latent field looks "almost" Gaussian, Rue *et al.* (2009). This is clearly assisted by the, non-negligible, impact of the Gaussian priors on the posteriors.

In our study, the prior distributions will be the default non informative and for the SPDE model, for σ and ϕ, we consider the Penalized Complexity prior, PC-prior, as derived in Fuglstad *et al.* (2018).

Results of parameter estimation

The results of the mean and standard errors of each parameter, obtained from a total of 500 independent samples are summarized in Table 4.2.

In Figures 4.7, 4.8 and 4.9 we have the corresponding boxplots for the preferential ($\beta = 2$), non-preferential ($\beta = 0$) and preferential ($\beta = -2$) simulated data sets, respectively, with true parameter values marked as red line. (PS corresponds to LAP method)

By analysing Table 4.2 and Figures 4.7, 4.8 and 4.9, we conclude that under Preferential Sampling, LAP, via TMB offers more accurate estimates than LAP via INLA, except in the case of σ. Comparing with the traditional Kalman filter, LAP showed considerable success mainly for μ, σ and ϕ. The parameter β seems to be underestimated using LAP and R-INLA in the case of $\beta = 2$ and overestimated for $\beta = -2$.

		PS Data $\beta = 2$			Not PS Data $\beta = 0$		
	True	LAP	INLA	CTS	LAP	INLA	CTS
$\widehat{\mu}$	**0**	**0.167** (0.483)	**0.600** (0.423)	**1.929** (0.480)	**-0.010** (0.386)	**-0.013** (0.381)	**0.003** (0.362)
$\widehat{\sigma}$	**1.581**	**1.471** (0.355)	**1.550** (0.333)	**0.906** (0.157)	**1.496** (0.201)	**1.580** (0.194)	**1.529** (0.207)
$\widehat{\phi}$	**5**	**5.873** (2.531)	**7.486** (3.097)	**2.339** (1.462)	**5.061** (1.605)	**5.536** (1.533)	**5.065** (1.672)
$\widehat{\tau}$	**0.1**	**0.166** (0.099)	**0.151** (0.352)	**0.176** (0.090)	**0.211** (0.144)	**0.045** (0.110)	**0.233** (0.135)
$\widehat{\beta}$	**2 ; 0**	**1.359** (0.258)	**1.076** (0.204)		**-0.005** (0.098)	**-0.004** (0.077)	

		PS Data $\beta = -2$		
	True	LAP	INLA	CTS
$\widehat{\mu}$	**0**	**-0.174** (0.384)	**-1.768** (1.880)	**-1.919** (0.480)
$\widehat{\sigma}$	**1.581**	**1.426**(0.279)	**1.699**(0.455)	**0.913** (0.153)
$\widehat{\phi}$	**5**	**5.223** (1.695)	**6.443**(1.748)	**2.317**(1.228)
$\widehat{\tau}$	**0.1**	**0.155**(0.101)	**0.080** (0.113)	**0.170**(0.094)
$\widehat{\beta}$	**-2**	**-1.344** (0.241)	**-0.530**(0.786)	

Table 4.2: Maximum likelihood estimates, under LAP (implemented via TMB package), LAP (implemented via INLA package) and by Kalman filter approach (implemented via cts package), mean (standard errors) obtained from a total of 500 independent samples.

For the case of non-preferential sampling, all estimation methods perform equivalent. The exception is in the case of τ, using LAP via INLA which presents an unexpected behavior (Figure 4.8).

Sensitivity Analysis

To investigate the sensitivity to initial values in parameter estimation, we conducted two different parameter estimations, one considering as initial values ($\boldsymbol{\theta_0}$) the "true" values and other considering as $\boldsymbol{\theta_0}$ the parameters estimated by traditional Kalman filter approach.

An estimate for the initial value of β, given a sample data set Y, can be obtained as follows. Suppose that $Y = \{(t_i, y_i) : i = 1, \ldots, n\}$, where y_i denotes the measured value and t_i is the corresponding time of the observation. A preliminary β_0 can be obtained through a simple algorithm such as: first, use a kernel-type intensity estimator of the locations to derive $\widehat{\lambda}(t)$; and, then, choose β_0 such that $\log \widehat{\lambda}(t) \simeq const + \beta_0 Y(t)$

The results of the mean and standard errors of each parameter, obtained from a total of 250 independent samples are summarized in Table 4.3.

The proposed method seems to be quite robust to initial values of θ in both scenarios, under preferential and not preferential sample data.

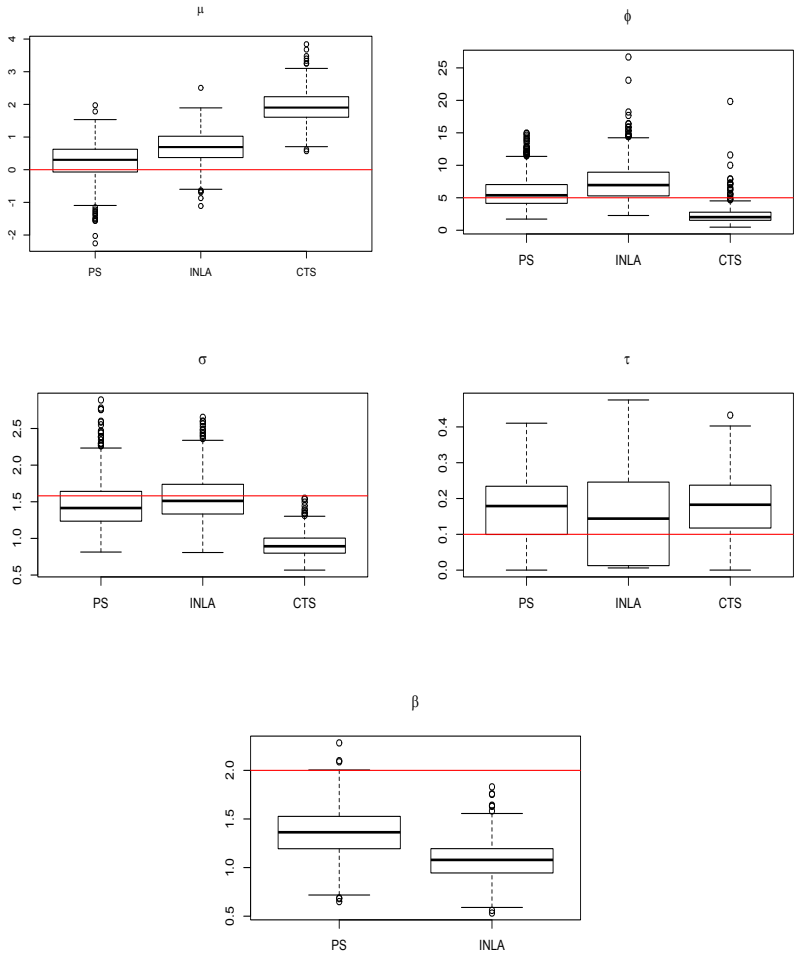


Figure 4.7: Boxplots for models parameters estimated over 500 preferentially sample simulated data sets, $\boldsymbol{\beta} = \mathbf{2}$, with true parameter values marked as red line.

4.5.3 Application to real data

In this Section, we apply our modelling approach supported by the LAP estimation method to the real data related to the biomedical marker, platelet, after a cancer patient undergoes a bone marrow transplant, previously described in Section 4.4.3.

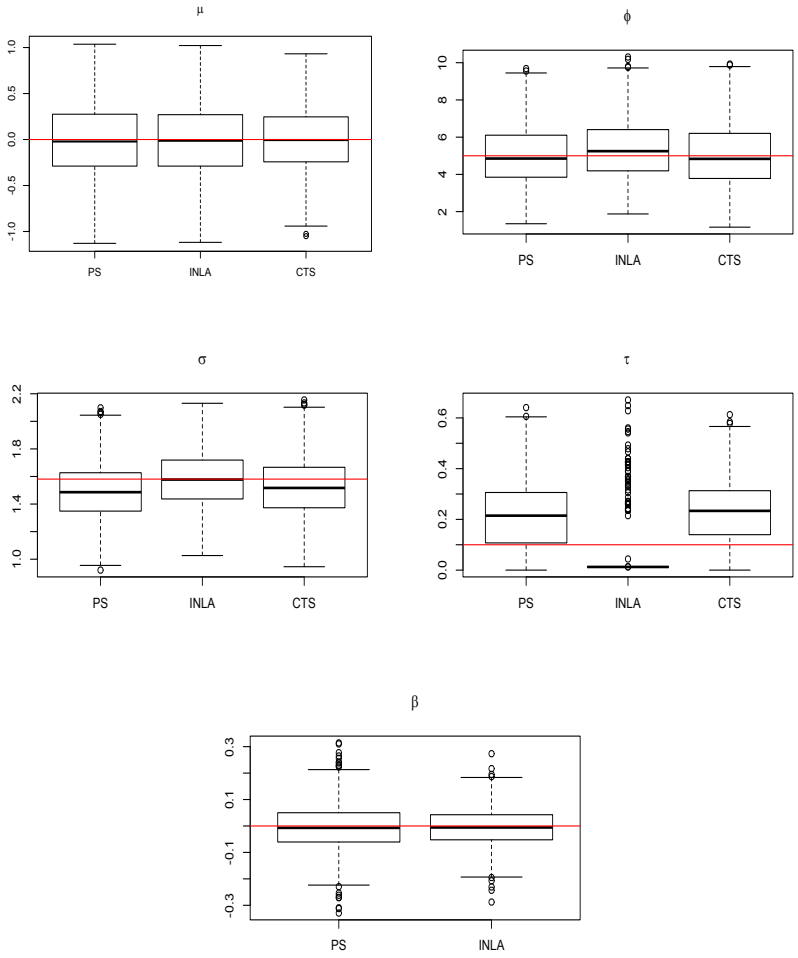


Figure 4.8: Boxplots for models parameters estimated over 500 non-preferentially sample simulated data sets, $\boldsymbol{\beta = 0}$, with true parameter values marked as red line.

Biomedical marker

The estimated parameters, using LAP method, together with estimated standard errors are summarized in Table 4.4.

Comparing the above parameter estimates with those obtained in Section 4.4.3, via Monte Carlo method: $\hat{\mu} = 4.97$, $\hat{\phi} = 54.85$, $\hat{\sigma} = 0.52$, $\hat{\tau} = 0.14$ and $\hat{\beta} = -1.51$, we conclude that the estimated value for β also has negative sign but a bit lower. Anyway, the corresponding confidence interval for $\widehat{\beta}$ is $(-1.568; -0.304)$, confirming

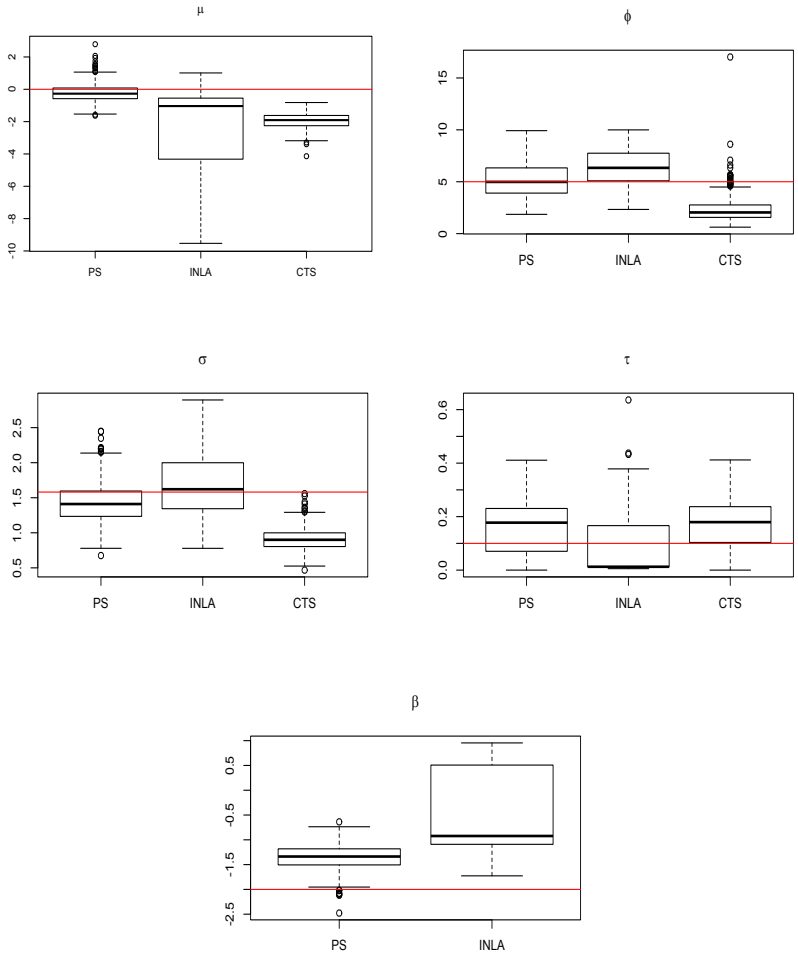


Figure 4.9: Boxplots for models parameters estimated over 500 preferentially sample simulated data sets, $\boldsymbol{\beta} = \mathbf{-2}$, with true parameter values marked as red line.

that β estimated from Monte Carlo and LAP approaches are in accordance. The estimates for the mean parameter, considering the two approaches, present equivalent results.

Analogous to what was done in Section 4.4.3, we plugging-in the estimated parameters in equations (4.8) and (4.9) and we obtain the predictions of the biomarker within the period of observations. Figure 4.10, top panel, shows the 95% prediction intervals for (log of) the biomarker, obtained from MCMLE's while the middle panel represents

		Preferential Data set		Not Preferential Data set	
	True	LAP (True $\boldsymbol{\theta_0}$)	LAP ($\boldsymbol{\theta_0}$ from CTS)	LAP (True $\boldsymbol{\theta_0}$)	LAP ($\boldsymbol{\theta_0}$ from CTS)
$\widehat{\mu}$	**0**	**0.247** (0.417)	**0.247** (0.417)	**-0.030** (0.388)	**-0.030** (0.388)
$\widehat{\sigma}$	**1.581**	**1.412** (0.255)	**1.413** (0.255)	**1.500** (0.207)	**1.500** (0.207)
$\widehat{\phi}$	**5**	**5.244** (1.699)	**5.242** (1.700)	**5.167** (1.739)	**5.167** (1.739)
$\widehat{\tau}$	**0.1**	**0.164** (0.104)	**0.163** (0.106)	**0.204** (0.138)	**0.203** (0.138)
$\widehat{\beta}$	**1.5;0**	**1.175** (0.159)	**1.175** (0.159)	**0.002** (0.100)	**0.002** (0.100)

Table 4.3: MLE's, mean (standard errors) obtained from a total of 250 independent samples, considering as initial values ($\boldsymbol{\theta_0}$) the "true" values and other considering the parameters estimated by traditional Kalman filter.

Parameter	Estimate	Standard Error
$\widehat{\mu}$	4.993	0.290
$\log(\widehat{\omega})$	2.545	0.198
$\widehat{\sigma}$	0.329	
$\log(\widehat{\phi})$	3.559	0.710
$\widehat{\phi}$	35.115	
$\log(\widehat{\tau})$	- 2.086	0.132
$\widehat{\tau}$	0.124	
$\widehat{\beta}$	-0.936	0.316

Table 4.4: Maximum likelihood estimates under LAP.

the 95% prediction intervals obtained from the MLE's from the Kalman filter approach and bottom panel represents the 95% prediction intervals obtained from the MLE's from LAP approach. In this situation the predictions obtained from LAP present lower variance than the predictions obtained from Monte Carlo approach, revealing greater precision.

4.6 Conclusions

We propose, in this Chapter, a methodology that takes into account the times of occurrence of the observations but also able to deal with irregularly spaced time series. Firstly we propose a Monte Carlo approach for the maximum likelihood estimation that not only provides good estimates for model parameters but also reveals quite satisfactory results for prediction. However this approach also presents some drawbacks and later in this Chapter we present a numerical alternative to the Monte Carlo Simulation.

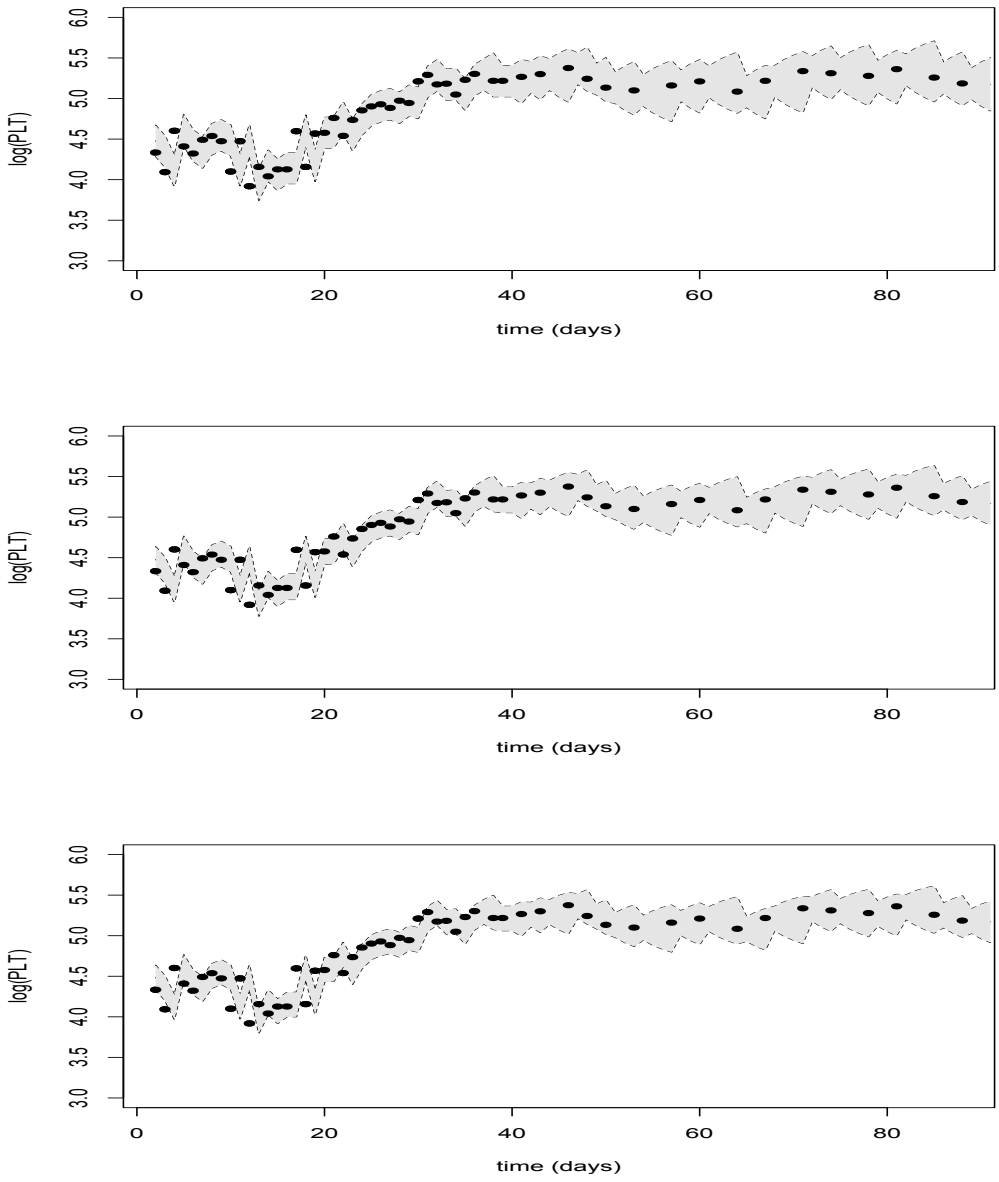


Figure 4.10: Prediction 95% confidence intervals using predictions acquired from MCMLE's (top), MLE's (middle) and LAP (bottom).

This approach, based on a Laplace approximation, increases the stability of our parameter estimates and presents quite satisfactory results for parameter estimation and predictions. It is much more computationally efficient and runs faster, while MCMLE's

takes approximately 20 minutes to estimate parameters in a single simulation, LAP takes approximately 21 seconds. Although INLA is slightly faster (16 seconds), LAP presents more accurate results and provides user high levels of flexibility, due to the direct specification of the joint likelihood.

A key aspect of this methodology is that it provides a tool to acquire some information on the underlying stochastic process.

In the next Chapter, we consider that the sample design may depend on all past observation times and actual observed values and we propose a model that allows to take into account the *history* of the process.

5

Modelling informative time points: an evolutionary process approach

Previously we assumed that the variable of interest is sampled in time according to a sampling design that depends on the values of the underlying process, ignoring the past of the observation processes. However, this kind of assumption of a memoryless process for the observations process having an evolution without aftereffects is sometimes unrealistic and useless in real contexts, where the dependence on the past is crucial.

In this Chapter we consider that the sampling design may depend on entire past history of the process, meaning all the times of the observations as well as the values of these observations. In these situations, the observed time points can be considered informative to the process being studied. The importance of joint modelling informative times and data was already recognised by Ryu *et al.* (2007) and Liang *et al.* (2009), within the scope of longitudinal studies. In current work, taking into account the natural temporal order underlying available data represented by a time series, then a modelling approach based on evolutionary processes seems a natural choice.

5.1 Introduction

Point processes provide, as noted before, a very useful theoretical tool to represent the evolution of some random value, or system, over time. In such processes it is assumed

that what happens now may depend on the past, but not on the future. This identifies a natural ordering for temporal point processes. Our interest now is to consider a point process that specifies a stochastic model for the time of the next event given we know all the times of previous events. Such processes are termed evolutionary point process.

In the next Section, a brief introduction to the theory of evolutionary processes, based on Daley & Vere-Jones (2003), is developed.

5.2 Evolutionary point processes

An important concept in evolutionary processes is the history of the process, denoted by H_t which represents the entire history of the point process prior to time t, meaning that H_t specifies the times of all point events in the interval $(-\infty, t)$. We refer to $\widetilde{H}_t$ as the observed history of the process over the interval $[0, t)$, that is the history consistent with an observation on the process. In this work, the specification of the point process conditional on its history is via the conditional intensity function, defined formely below. Furthermore, the point processes are assumed to be simple point processes, meaning that no points coincide and therefore the points can be ordered strictly in time.

5.2.1 Conditional intensity function

The conditional intensity function, $\lambda^*(t) = \lambda(t|\widetilde{H}_t)$, is defined by

$$\lambda^*(t) = \frac{f(t|\widetilde{H}_t)}{1 - F(t|\widetilde{H}_t)},\ t_1 < \cdots < t_n < t \tag{5.1}$$

where $f(t|\widetilde{H}_t)$ is the conditional density and $F(t|\widetilde{H}_t)$ is the corresponding cumulative distribution function and $t_1, \cdots, t_n$ are observed points.

Intuitively, the conditional intensity at t gives the conditional "risk" of a point event occurring at that instant in time, given the observed history of the process prior to time t.

Examples of point processes in which the conditional intensity has a particular functional form are the following:

- The (inhomogeneous) Poisson process. In this process the number of points in disjoint sets is independent and the conditional intensity function inherets this property. The Poisson process is quite simply the point process in which the

conditional intensity function is independent of the past, i.e. the conditional intensity function is equal to the intensity function of the Poisson process, $\lambda^*(t) = \lambda(t)$.

- The conditional intensity function of a Hawkes process with an exponential decay function has the form

$$\lambda^*(t) = \eta + \psi \sum_{i:t_i \in (0,t)} \exp(-\gamma(t - t_i))$$

where $\eta > 0$, $\psi \geq 0$, $\gamma > 0$ and $\psi < \gamma$ if the process is assumed to be stationary. Note that each time a new point arrives in this process, the conditional intensity grows by ψ and then decreases exponentially back towards η. In other words, a point increases the chance of getting other points immediately after (self-exciting). Setting $\psi = 0$, return us to the homogeneous Poisson process.

5.2.2 Marked point processes

In addition to the times of the point events, there may be additional variables that are of interest associated with each point event. This information is known as marks and the mark space (M) can be of many different types, but it is often (a subset of) $\mathbb{R}$ or $\mathbb{N}$. The marks may have an independent interest or may be included to make a more realistic model of the event times. For example, in the analysis of a particular medical indicator, it is relevant to know its value and not only when it was observed. In addition, the value of the indicator influences how often measurements are taken.

More formally, a marked point process, with point event times in $\mathbb{R}$ and marks in M, is a point process $\{(t_i, y_i)\}$ on $\mathbb{R} \times M$ with the additional property that the process associated with times $t_1, t_2, \cdots$, the ground process, is itself a point process on $\mathbb{R}$. We specify a marked point process by defining the conditional intensity $\lambda(\cdot|\widetilde{H}_t)$ of the ground process, and then, for a given point event and observed history at time t, we define the conditional distribution function for the marks. We can specify the distribution of the mark y associated with the point t by its conditional density function $f^*(y|t) = f(y|t, \widetilde{H}_t)$, i.e. this specifies the distribution of the mark given t and the history of the process, that includes information of times and marks of past events. The definitions of the complete and observed histories, H_t, and $\widetilde{H}_t$, and the conditional

intensity function are extended for marked point processes. We can now define the conditional intensity function for the marked case as

$$\lambda^*(t, y) = \lambda^*(t) f^*(y|t) \tag{5.2}$$

where $\lambda^*(t)$, called the ground intensity, is the counting process associated with the point events in the time domain and is defined exactly as the conditional intensity function for the unmarked case, except that it is allowed to depend on the marks of the past events. In addition, the marks are assumed to be conditionally independent given the history of the marked point process and unpredictable. A process is said to have unpredictable marks if the distribution of the mark at t_i is independent of all previous point event times and marks.

Thus, we can rewrite (5.2) as

$$\lambda^*(t, y) = \frac{f(t|\widetilde{H}_t) f^*(y|t)}{1 - F(t|\widetilde{H}_t)} = \frac{f(t, y|\widetilde{H}_t)}{1 - F(t|\widetilde{H}_t)}$$

where $f(t, y|\widetilde{H}_t)$ is the joint density of the time and the mark, conditional on past times and marks, and $F(t|\widetilde{H}_t)$ is the conditional cumulative distribution function of t also conditional on the past times and marks.

An example of a marked point process is the marked Hawkes process. This process is a generalization of the unmarked Hawkes process, such that each point event time now has a mark associated with it. The conditional intensity of the ground process is given by

$$\lambda(t|\widetilde{H}_t) = \lambda^*(t) = \eta + \psi \sum_{t_i : t_i \in (0,t)} \exp(\beta_1 y_i) \exp(-\gamma(t - t_i)) \tag{5.3}$$

where $\eta, \gamma > 0$, $\psi, \beta_1 \geq 0$ and y_i denotes the observed value at time t_i.

Equivalently we could define it by its conditional intensity function including both marks and times

$$\lambda^*(t, y) = \left(\eta + \psi \sum_{t_i : t_i \in (0,t)} \exp(\beta_1 y_i) \exp(-\gamma(t - t_i)) \right) f^*(y|t) \tag{5.4}$$

The idea behind using this model is that every new event increases the intensity by $\psi \exp(\beta_1 y_i)$ and large events increase the intensity more than small.

5.2.3 Inference

Daley & Vere-Jones (2003) note that for point processes described as having an evolutionary character, their conditional intensities and likelihoods are relatively simple. The evolutionary character of such point processes allows the likelihood to be found by successively conditioning on the past. Explicitly, the likelihood of a realization $((t_1, y_1), \cdots, (t_n, y_n))$ on $[0, T) \times \mathbb{R}$, of a marked point process is given by

$$L_E = \left(\prod_{i=1}^{n} \lambda^*(t_i) \right) \exp\left(- \int_0^T \lambda^*(u) du \right) \left(\prod_{i=1}^{n} f^*(y_i|t_i) \right) \tag{5.5}$$

See (Daley & Vere-Jones, 2003, p.246-256) for a development of the likelihood. The third factor on the right-hand side of (5.5) is the contribution to the likelihood from the observed marks. The associated log-likelihood function is given by

$$\log(L_E) = \sum_{i=1}^{n} \log \lambda^*(t_i) - \int_0^T \lambda^*(u) du + \sum_{i=1}^{n} \log f^*(y_i|t_i) \tag{5.6}$$

The use of the log-likelihood implies bearing in mind some practical considerations. A point process is only observed for a finite interval $[0, T]$ and time 0 is some time after the origin of the process. For evolutionary point processes, there may be effects from point events occurring before time 0. Daley & Vere-Jones (2003) referred such effects as edge or boundary effects. An approach often taken in the literature is ignoring the effects from point events occurring before the start of the observation period. In this case the conditional intensity can be regarded as approximate for some initial part of the observation period, and as such, there is likely to be some effect on the estimated model. Rasmussen (2013) highlights that the estimate of η is likely to be too high, however, he noted that the effects on the estimated model will be negligible if the data set being used is large. Another question is the computational burden of evaluating (5.6), this arises from the nested sum in the first term, if we take into account that

$$\sum_{i=1}^{n} \log \lambda^*(t_i) = \sum_{i=1}^{n} \log \left(\eta + \psi \sum_{t_i : t_i \in (0,t)} \exp(\beta_1 y_i) \exp(-\gamma(t - t_i)) \right) \tag{5.7}$$

5.3 An evolutionary model for informative time points

Consider an unobserved stochastic process in time $S(t)$, represented by a CAR(1). Now admit that $S(t)$ is observed at times t_i, $i = 1, \cdots, n$, yielding a data set (t_i, y_i), where the corresponding $Y_i = Y(t_i)$ is the noisy version of $S(t_i)$. Since our goal is to inference on $S(t)$, admitting that the sampling times are stochastic and the sampling design may depend on all past history of the process, (both the actual times and values of the observations), then a model able to deal with this evolutionary character must specify the joint distribution of S, $T = (t_1, \ldots, t_n)$ and $Y = (Y_1, \ldots, Y_n)$, $[S, T, Y]$. Considering that $[S, T, Y] = [S][T, Y\,|S]$ let $\{(T, Y)|S\}$ be an evolutionary marked point process with ground intensity

$$\lambda_S^*(t) = \lambda(t|\widetilde{H}_t, S) = \eta + \psi \sum_{t_i : t_i \in (0,t)} \exp(\beta_1 y_i) \exp(-\gamma(t - t_i)) \tag{5.8}$$

with $\eta, \gamma > 0$ and $\psi, \beta_1 \geq 0$.

Admitting the conditional mark density, $f_S^*(y|t) = f^*(y|t, S)$, then according to (5.2), the conditional intensity function including both marks and times is

$$\lambda_S^*(t, y) = \lambda(t, y|\widetilde{H}_t, S) = \lambda_S^*(t) f_S^*(y|t) \tag{5.9}$$

The main purposes behind this model are

- every new event increases the intensity by $\psi \exp(\beta_1 y_i)$ and large events increase the intensity more than small events;
- observations that are more distant in time have less influence, considering on γ parameter;
- the initial value of the conditional intensity equals η and we ignore effects from events occurring before the first observation.

5.3.1 Maximum likelihood estimation

To obtain estimates for the parameters of the model we use maximum likelihood estimation. For the shared latent process model, the likelihood function for data T and Y can be expressed as

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [T, Y, S] dS = \int_S [S][T, Y|S] dS \tag{5.10}$$

where $\boldsymbol{\theta} = (\mu, \sigma_w, \alpha_0, \tau, \beta_1, \gamma, \psi, \eta)$ represents all the model parameters.

Considering that the likelihood of a marked point process is given by (5.5), $[T, Y|S]$ in (5.10) takes the form

$$[T, Y|S] = \left(\prod_{i=1}^{n} \lambda_S^*(t_i) \right) \exp\left(- \int_0^T \lambda_S^*(u) du \right) \left(\prod_{i=1}^{n} f_S^*(y_i|t_i) \right)$$

The associated log-likelihood function is given by

$$\log([T, Y|S]) = \sum_{i=1}^{n} \log \lambda_S^*(t_i) - \int_0^T \lambda_S^*(u) du + \sum_{i=1}^{n} \log f_S^*(y_i|t_i) \tag{5.11}$$

Substituting in (5.11), the conditional (ground) intensity, $\lambda_S^*(\cdot)$, and the conditional mark density $f_S^*(y_i|t_i)$, specified as $N(S_i, \tau^2)$, then the log-likelihood can be rewritten as

$$\begin{aligned} \log([T, Y|S]) \quad = \quad & \sum_{i=1}^{n} \log \left(\eta + \psi \sum_{j: t_j < t_i \in (0,t)} \exp(\beta_1 y_j - \gamma(t_i - t_j)) \right) \\ & -\eta T - \frac{\psi}{\gamma} \sum_{i=1}^{n} \exp(\beta_1 y_i)(1 - \exp(-\gamma(T - t_i))) \\ & -\frac{n}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \sum_{i=1}^{n} (y_i - S_i)^2 \end{aligned} \tag{5.12}$$

5.3.2 Computational issues

To overcome the computational burden of evaluating (5.12), which arises from the nested sum in the first term

$$\sum_{i=1}^{n} \log \left(\eta + \psi \sum_{j:t_j<t_i \in (0,t)} \exp(\beta_1 y_j - \gamma(t_i - t_j)) \right)$$

we use a compiled C++ subroutine. Besides that we set $\lambda(0|\widetilde{H}_0) = \eta$ and ignore the effects from point events occurring before time 0.

For $[S]$ in (5.10) we adopt SPDE approximation for the Gaussian process S, as described in Section 4.5.1. This allows to create a temporal mesh and corresponding components of the sparse precision matrix of a GMRF, used to more efficiently evaluate the normal density required.

To approximate the integral in the likelihood (5.10) we utilize Automatic Differentiation of a Laplace Approximation to the marginal likelihood, as described in Section 4.5.1. In the case of sampling design that may depend on entire past history of the process, we simply have to define the joint negative log-likelihood as

$$f(S, \boldsymbol{\theta}) = -log([S][T, Y|S])$$

and allow TMB package to integrate out the latent field S to evaluate approximately (5.10).

This model, henceforth EVOL, allows to take into account the history of the process, capture the evolutionary character of the process and deal with irregularly spaced time series.

5.4 Numerical studies

We now intend to proceed with the assessment of the EVOL model, comparing the results of its parameter estimates and those of the traditional Kalman filter approach. We use simulated time series, so we start by describing the procedure needed to simulate a marked point process.

Simulation Design

The method to simulate an inhomogeneous Poisson process, described in Section 4.2.3 requires that the conditional intensity to be bounded above, i.e. there is a finite M such that for all t, $\lambda(t|\widetilde{H}_t, S) \leq M$. This method was generalised by Ogata (1981) and this generalisation only requires that the intensity to be locally bounded. The algorithm is described as follows. Suppose we can find a piecewise constant process $M(\cdot|\widetilde{H}_t, S)$, conditional on the history of the point process, such that for $t \in [0, T)$,

$$\lambda(t|\widetilde{H}_t, S) \leq M(\cdot|\widetilde{H}_t, S)$$

Given that we can find a suitable $M(\cdot|\widetilde{H}_t, S)$, we can simulate a realisation of the point process of interest in this way: define an inhomogeneous Poisson process N^* which has a piecewise constant intensity $M(\cdot|\widetilde{H}_t, S)$ that changes value according to the history $\widetilde{H}_t$ and decide on the termination condition, for e.g. the simulation interval is $[0, T)$, then simulate the points $0 \leq t_1^* < t_2^* < \cdots < t_{N^*[0,T)}^* < T$ from the process N^*. Each t_i^* is then selected with probability $\lambda(t_i^*|\widetilde{H}_{t_i^*}, S_{t_i^*})/M(t_i^*|\widetilde{H}_{t_i^*}, S_{t_i^*})$ to form part of the simulated realisation of the point process of interest, where the history $H_{t_i^*}$ and $S_{t_i^*}$ give the simulated history of the point process of interest up to time t_i^*. For each point t_i that is selected to the simulated realisation of the point process of interest we simulate a mark y_i from $Y(t) = \mu + S(t) + N(0, \tau^2)$.

In practice, the function $M(\cdot|\widetilde{H}_t, S)$ changes value each time a point event is added to the simulated realisation of the process of interest, and so it will not be known before carrying out the simulation.

To generate a time series under a Preferential Sampling that depends on all past history of the process, we adapt the R code used by (Lapham, 2014, p.124-125).

As follows, we start to generate a realization of S, a CAR(1) process with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$. These values correspond to $Var[S(\cdot)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ and $\phi = \frac{1}{\alpha_0} = 5$. The parameter values used to generate the marked point process are

$$\eta = 0.05, \psi = 0.025, \beta_1 = 0.6, \gamma = 0.1$$

and to generate the marks y_i, we consider $\mu = 0$ and $\tau = 0.1$.

To illustrate the results of these sampling procedure, we represent in Figure 5.1 a realization of the process S (gray line) and the resulting data set.

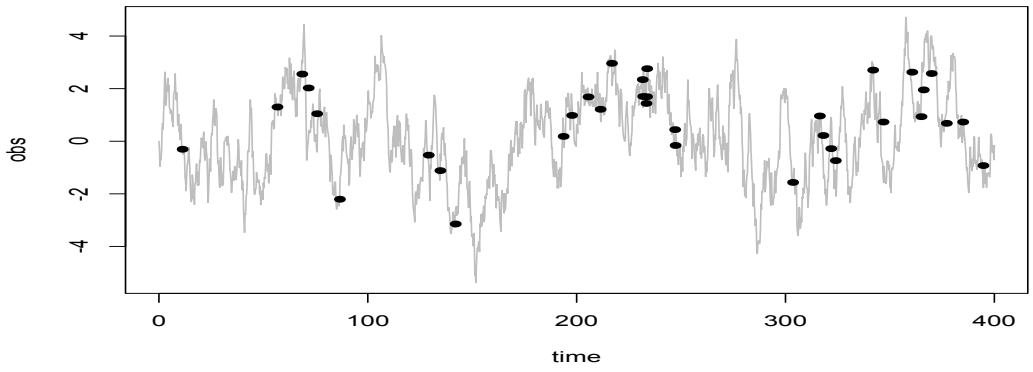


Figure 5.1: Sample times with dependency on all past history of the process and underlying process S (gray line).

Estimation Results

For EVOL model, η and ψ parameters have a tuning role. Relatively to η, we ignore effects from point events occurring before the start of the observation period and we assume that the initial value of the conditional intensity equals η. Regarding ψ, it controls the sum value in the ground intensity. Thus, in a first simulation study the parameters μ, σ, ϕ, τ, β_1 and γ are the target of estimation and we set η and ψ values at the true ones. For the simulation study we consider a total of 500 independent samples with at least 50 points over the interval $[0, 400]$. The results of the mean and the standard errors for each parameter, obtained from EVOL model, under (5.11), and from Kalman filter approach implemented via cts package are summarized in Table 5.1. In Figure 5.2 we have the corresponding boxplots, with true parameter values marked as red line.

By analysing Table 5.1 and Figure 5.2, we conclude that EVOL model presents more accurate estimates than Kalman filter approach. The parameter τ seems to be overestimated in both approaches. For β_1 and γ the estimates are quite reasonable and

	True	EVOL	CTS
$\widehat{\mu}$	**0**	**0.196** (0.267)	**0.225** (0.304)
$\widehat{\sigma}$	**1.581**	**1.567** (0.204)	**1.606** (0.209)
$\widehat{\phi}$	**5**	**5.995** (1.647)	**6.188** (1.617)
$\widehat{\tau}$	**0.1**	**0.456** (0.197)	**0.483** (0.194)
$\widehat{\beta 1}$	**0.6**	**0.618** (0.128)	
$\widehat{\gamma}$	**0.1**	**0.095** (0.026)	

Table 5.1: Maximum likelihood estimates, under EVOL approach and by Kalman filter approach, mean (standard errors) obtained from a total of 500 independent samples.

we believe that the inclusion of these two parameters in the model is more realistic in real contexts.

Further studies with different combinations of the parameters, namely for β_1 and γ were analysed. When $\beta_1 > \gamma$ the conclusions are similar, but when $\beta_1 < \gamma$ or $\beta_1 > 1$ it is necessary to do some calibration work with parameter ψ in order to obtain samples with a reasonable dimension.

Sensitivity Analysis

To analyse the impact of estimating also the parameters η and ψ and to investigate the sensitivity to initial values in parameter estimation, we conducted a second simulation study. We consider two different parameter estimations, one considering as initial values ($\boldsymbol{\theta_0}$) the "true" values and other considering for μ, ϕ, σ and τ the parameters estimated by traditional Kalman filter approach and for the other parameters we consider $\beta_1 = 0.4, \gamma = 0.2, \eta = 0.07$ and $\psi = 0.035$.

The results of the mean and standard errors of each parameter, obtained from a total of 200 independent samples are summarized in Table 5.2.

The proposed method seems to be quite robust to initial values and the inclusion of parameters η and ψ seems do not cause identifiability issues, only parameter β_1 is a little overestimated.

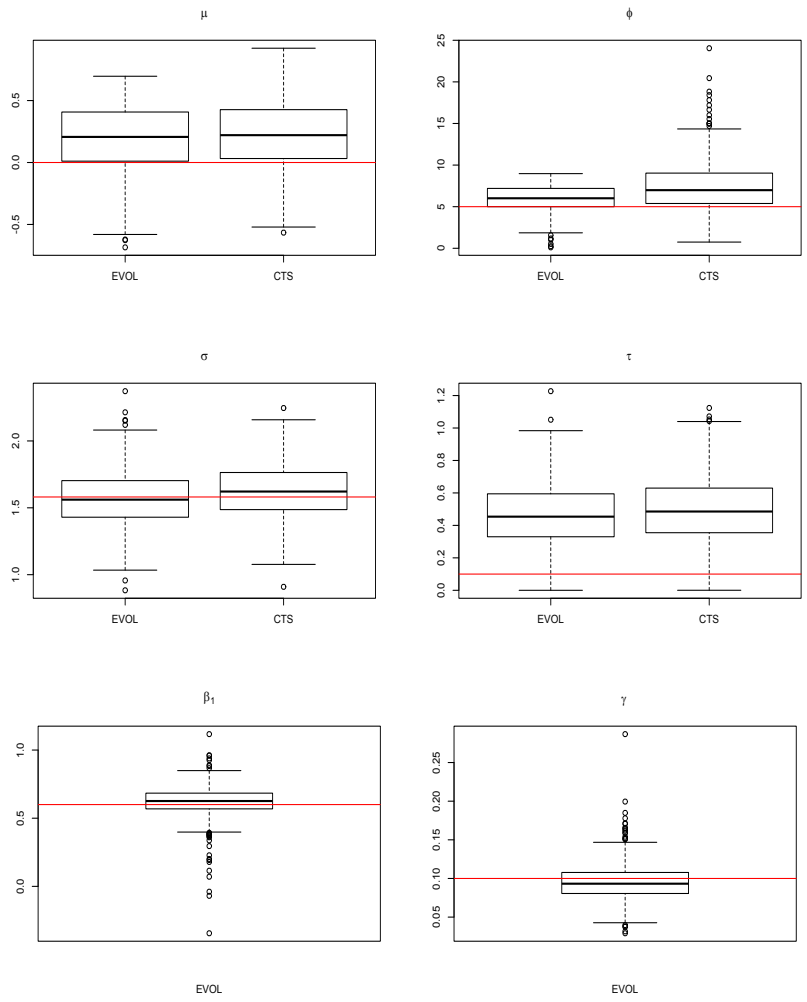


Figure 5.2: Boxplots for models parameters estimated over 500 independent samples with true parameter values marked as red line.

5.5 Conclusions

In this Chapter, we present a model approach that allows to deal with sampling designs that depend on all past history of the process. This model allows to take into account the evolutionary character of the process and is, in our opinion more realistic, since it also consider the previous observations and the temporal distance to which they occurred. The results for the parameter estimation are quite satisfactory and the algorithm is computationally efficient.

	True	EVOL (True $\boldsymbol{\theta_0}$)	EVOL ($\boldsymbol{\theta_0}$ from Kalman filter)
$\widehat{\mu}$	**0**	**0.239** (0.276)	**0.239** (0.276)
$\widehat{\sigma}$	**1.581**	**1.526** (0.224)	**1.526** (0.224)
$\widehat{\phi}$	**5**	**5.720** (1.429)	**5.720** (1.429)
$\widehat{\tau}$	**0.1**	**0.477** (0.202)	**0.477** (0.202)
$\widehat{\beta 1}$	**0.6**	**0.782** (0.325)	**0.782** (0.325)
$\widehat{\gamma}$	**0.1**	**0.114** (0.065)	**0.114** (0.065)
$\widehat{\psi}$	**0.025**	**0.025** (0.023)	**0.025** (0.023)
$\widehat{\eta}$	**0.05**	**0.065** (0.021)	**0.065** (0.021)

Table 5.2: MLE's, mean (standard errors) obtained from a total of 200 independent samples, considering as initial values for EVOL approach the parameters estimated by traditional Kalman filter.

5. MODELLING INFORMATIVE TIME POINTS: AN EVOLUTIONARY PROCESS APPROACH

6

Concluding remarks and further work

The main objective of this work was centered in the presentation of contributions to Spatial and Temporal modelling. Namely, in the context of analysing data under irregular sampling and data where the process of observation times/locations is stochastic and provides additional information about the phenomena under study.

The framework proposed in Chapter 3 allows inference on the large-scale and small-scale variation components of the spatio-temporal stochastic process. Our proposal uses a block bootstrap procedure to correctly assess uncertainty in parameter estimates and produce reliable confidence regions for (space-time) unobserved values of the variable of interest.

Nonetheless, the discussed model presents some limitations, one of which is the difficulty in capturing temporal specificities intrinsic to a location. In fact, as discussed in section 3.4.3 in the illustrating example, although the overall mean intra-day pattern of the NO_2 concentrations is well described by the model, individual stations and days present particularities that remain unexplained. For example, stations located in the surroundings of major cities present anticipated and/or postponed rush-hour traffic leading to lagged peaks of NO_2 concentrations. To overcome this issue interactions between harmonic regression and type of station could be incorporated into the model, or time and space-varying model parameters could be allowed. Furthermore, this method, as a two-stage approach may introduce some extra-variance in the inferential proce-

dures, which is expected to be negligible. An alternative approach to model NO_2 data is to use a technique based on the stochastic partial differential equations (SPDE) implemented via the INLA R package which is currently widely used in spatio-temporal modelling. For high resolution time series, such as the ones considered in the present work, Blangiardo & Cameletti (2015) point out that INLA becomes computationally expensive and advise lowering the temporal resolution by defining the model on a set of time knots, instead of on the set of all the time points. In our view, this could, however, mask high frequency variability, such as intra-day variability resulting from anthropogenic activities and meteorological conditions.

In Chapter 4 we propose a model-based approach that takes into account the times of occurrence of the observations but also able to deal with irregularly spaced time series. For parameter estimation we use maximum likelihood estimation and we propose two alternatives, one based on Monte Carlo simulation and other based on a Laplace approach to optimize the likelihood. The results for estimated parameters, in both situations, are quite satisfactory when compared with the traditional approach that uses Kalman filter to deal with irregularly spaced time series. However, the second one is much more computationally efficient, runs faster and increases the stability of our parameter estimates.

Diggle & Giorgi (2017), in the context of spatial statistics, affirm that the use of a single parameter in (4.6), β, to capture both the strength of the preferentiality and the amount of non-uniformity in sampling locations is somewhat inflexible. These authors discuss a more flexible and computational more efficient class of models, based on the proposal of Pati *et al.* (2011). Furthermore they suggest an extension to the model proposed by Diggle *et al.* (2010), by adding a second Gaussian process and use of stochastic partial differential equation models. For future investigation we intend to adapt those suggestions to the time dimension.

In Chapter 5 we present a model approach that allows to deal with sampling designs that depend on all past history of the process. To specify a process conditional on the past we considered the intensity function and a marked point process for the times T and marks Y. The suggested modelling approach exhibited, in the numerical studies, accurate estimates for the parameters and proved to be computationally quite efficient.

The new Protection Policy of Personal Data made it very difficult for us to find an illustrative data set for the framework presented in Chapter 5, but we intend to apply this model-based approach to a real data set and define a simple method to choose suitable starting values for the estimation algorithm.

As a goal for future research, we plan to investigate new model structures that accommodate covariates and allow for a non-gaussian response variable.

It is also of our interest to proceed with spatio-temporal modelling of data that presents both spatial and temporal Preferential Sampling. One such example rises in the context of *smart cities* projects which aim at helping city planners to correctly manage urban environments. This typically requires the development of statistical tools capable of handling large amounts of data collected by sensor networks. However, for practical constraints, the sampling design underlying the sensor networks might not uniformly represent the observation region, if more sensors are placed on those areas considered as more critical, leading to a Preferential Sampling design in space. In a similar way, the collection procedure along time might depend, also for practical constraints, on the observed values, if, for example, it is decided to monitor according the history of the process. An extension to spatio-temporal modelling of data may prove to be useful for the above examples.

References

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis, NY: Wiley*. 53

Belcher, J., Hampton, J. & Wilson, G.T. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 141–155. 11, 58

Bell, B.M. (2012). Cppad: a package for C++ algorithmic differentiation. *Computational Infrastructure for Operations Research*, **57**. 61

Besag, J. (1991). Rejoinder. *Annals of the institute of statistical mathematics*, **43**, 45–59. 7

Bivand, R.S., Pebesma, E. & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. 27

Blangiardo, M. & Cameletti, M. (2015). Spacetime model lowering the time resolution. In *Spatial and spatio-temporal Bayesian models with R-INLA*, chap. 8, 295–303, John Wiley & Sons. 88

Box, G.E., Jenkins, G.M., Reinsel, G.C. & Ljung, G.M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons. 43

Brockwell, P. (2001). Continuous-time ARMA processes. *Handbook of Statistics*, **19**, 249–276. 12

Brockwell, P.J. (2009). Lévy–driven continuous–time arma processes. *Handbook of financial time series*, 457–480. 11

REFERENCES

Brockwell, P.J. & Davis, R.A. (2002). *Introduction to time series and forecasting*. springer. 43, 50

Bruno, F., Guttorp, P., Sampson, P.D. & Cocchi, D. (2003). Non-separability of space-time covariance models in environmental studies. In *The ISI International Conference on Environmental Statistics and Health*, 141, 153, Univ Santiago de Compostela. 25

Cameletti, M., Ignaccolo, R. & Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, **22**, 985–996. 19

Carslaw, D.C. (2005). Evidence of an increasing NO2/NOx emissions ratio from road traffic emissions. *Atmospheric Environment*, **39**, 4793–4802. 18

Cressie, N. (1993). *Statistics for spatial data, Revised Edition*. Wiley, New-York. 6, 7, 9

Cressie, N. & Huang, H.C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1339. 25

Daley, D.J. & Vere-Jones, D. (2003). *An introduction to the theory of point processes. Volume I: Elementary Theory and Methods,*. Springer, New York, 2nd edn. 8, 44, 74, 77

Daley, D.J. & Vere-Jones, D. (2008). *An introduction to the theory of point processes. Volume II: general theory and structure*. Springer, New York, 2nd edn. 8

Daniels, M.J. & Hogan, J.W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press. 44

De Cesare, L., Myers, D. & Posa, D. (2001). Estimating and modelling space–time correlation structures. *Statistics & Probability Letters*, **51**, 9–14. 25

De Iaco, S. (2010). Space–time correlation analysis: a comparative study. *Journal of Applied Statistics*, **37**, 1027–1041. 25

De Iaco, S. & Posa, D. (2012). Predicting spatio-temporal random fields: some computational aspects. *Computers & Geosciences*, **41**, 12–24. 27, 30

De Livera, A.M., Hyndman, R.J. & Snyder, R.D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, **106**, 1513–1527. 18

Diggle, P. & Giorgi, E. (2017). Preferential sampling of exposure levels. In *Handbook of Environmental and Ecological Statistics*, chap. 21, CRC Press. 60, 88

Diggle, P. & Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied statistics*, 49–93. 44

Diggle, P.J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman and Hall/CRC. 8

Diggle, P.J. & Ribeiro, P.J. (2007). *Model-based Geostatistics*. Springer. 8

Diggle, P.J., Menezes, R. & Su, T.l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232. 1, 11, 44, 52, 88

Dinsdale, D. & Salibian-Barrera, M. (2018). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 60

Donnelly, A., Misstear, B. & Broderick, B. (2011). Application of nonparametric regression methods to study the relationship between NO2 concentrations and local wind direction and speed at background sites. *Science of the Total Environment*, **409**, 1134–1144. 18

Doob, J.L. (1953). *Stochastic processes*. Wiley New York. 14

Erdogan, E., Ma, S., Beygelzimer, A. & Rish, I. (2005). Statistical models for unequally spaced time series. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, 626–630, SIAM. 12

European Environment Agency, EEA (2015). Air quality in europe - 2015 report. 18

Fassò, A. & Negri, I. (2002). Non-linear statistical modelling of high frequency ground ozone data. *Environmetrics*, **13**, 225–241. 18

Fonseca, T.C. & Steel, M.F. (2011). A general class of nonseparable space–time covariance models. *Environmetrics*, **22**, 224–242. 25

Fuglstad, G.A., Simpson, D., Lindgren, F. & Rue, H. (2018). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 1–8. 65

Ghorbani, H., Möller, H. & Stoyan, D. (2006). Using pareto and weibull distributions in the modelling of growth processes: theory and methods. *South African Statistical Journal*, **40**, 75–98. 46

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, **97**, 590–600. 25

Grice, S., Stedman, J., Kent, A., Hobson, M., Norris, J., Abbott, J. & Cooke, S. (2009). Recent trends and projections of primary NO2 emissions in europe. *Atmospheric Environment*, **43**, 2154–2167. 18

Griewank, A. & Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*, vol. 105. Siam. 61

Hogan, J.W. & Laird, N.M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in medicine*, **16**, 259–272. 44

Hyndman, R.J. (1992). *Continuous time threshold autoregressive models*. Ph.D. thesis, University of Melbourne. 13

Hyndman, R.J. & Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, **22**, 679–688. 32

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, vol. 70. John Wiley & Sons. 8

Ip, R.H. & Li, W. (2015). Time varying spatio-temporal covariance models. *Spatial Statistics*, **14**, 269–285. 25

Jones, R.H. (1981). Fitting a continuous time autoregression to discrete data. In *Applied time series analysis II*, 651–682, Elsevier. 11, 12, 15

Jones, R.H. (1984). Fitting multivariate models to unequally spaced data. In *Time series analysis of irregularly observed data*, 158–188, Springer. 57

Jones, R.H. (1985). Time series analysis with unequally spaced data. *Handbook of statistics*, **5**, 157–177. 11

Kelly, B.C., Becker, A.C., Sobolewska, M., Siemiginowska, A. & Uttley, P. (2014). Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets. *The Astrophysical Journal*, **788**, 33. 12

Keramatinia, A., Hassanipour, S., Nazarzadeh, M., Wurtz, M., Monfared, A.B., Khayyamzadeh, M., Bidel, Z., Mhrvar, N. & Mosavi-Jarrahi, A. (2016). Correlation between nitrogen dioxide as an air pollution indicator and breast cancer: a systematic review and meta-analysis. *Asian Pacific Journal of Cancer Prevention*, **17**, 419–424. 17

Kreiss, J.P. & Paparoditis, E. (2011). Rejoinder: Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, **40**, 393–395. 26

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B.M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21. 61

Lai, H.K., Tsang, H. & Wong, C.M. (2013). Meta-analysis of adverse health effects due to air pollution in chinese populations. *BMC Public Health*, **13**, 360. 17

Lange, K. (2010). *Numerical analysis for statisticians*. Springer Science & Business Media. 12

Lapham, B.M. (2014). *Hawkes processes and some financial applications*. Ph.D. thesis, University of Cape Town. 81

Li, Z. (2014). *Methods for Irregularly Sampled Continuous Time Processes*. Ph.D. thesis, UCL (University College London). 11

Liang, Y., Lu, W. & Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*, **65**, 377–384. 44, 73

Lin, H., Scharfstein, D.O. & Rosenheck, R.A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 791–813. 44

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498. 63, 64

Little, R.J. & Rubin, D.B. (2014). *Statistical analysis with missing data*. John Wiley & Sons. 12

Ma, C. (2008). Recent developments on the construction of spatio-temporal covariance models. *Stochastic Environmental Research and Risk Assessment*, **22**, 39–47. 25

McCarthy, M.C., O'Brien, T.E., Charrier, J.G. & Hafner, H.R. (2009). Characterization of the chronic risk and hazard of hazardous air pollutants in the united states using ambient monitoring data. *Environmental health perspectives*, **117**, 790. 17

Menezes, R., Piairo, H., García-Soidán, P. & Sousa, I. (2016). Spatial–temporal modellization of the NO2 concentration data through geostatistical tools. *Statistical Methods & Applications*, **25**, 107–124. 18, 19, 20

Møller, J. & Waagepetersen, R.P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC. 8

Møller, J., Syversveen, A.R. & Waagepetersen, R.P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, **25**, 451–482. 47

Monteiro, A., Menezes, R. & Silva, M.E. (2017). Modelling spatio-temporal data with multiple seasonalities: the NO2 portuguese case. *Spatial Statistics*, **22**, 371–387. 17

Myers, D.E. (2004). Estimating and modelling space-time variograms. In *Proceedings of the joint meeting of TIES-2004 and ACCURACY-2004*. 27

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, **27**, 23–31. 81

Pati, D., Reich, B.J. & Dunson, D.B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**, 35–48. 88

Pebesma, E.J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, **30**, 683–691. 27

Pinheiro, J.C. & Bates, D.M. (2001). *Mixed-Effects Models in S and S-plus*. Springer New York. 15

Priestley, M.B. (1981). Spectral analysis and time series. 12

Qualar (2015). Online database on air quality. 19

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 27

Rasmussen, J.G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, **15**, 623–642. 77

Rathbun, S.L. (1996). Estimation of poisson intensity using partially observed concomitant variables. *Biometrics*, 226–242. 47

Ricciardolo, F.L., Sterk, P.J., Gaston, B. & Folkerts, G. (2004). Nitric oxide in health and disease of the respiratory system. *Physiological reviews*, **84**, 731–765. 18

Roberts-Semple, D., Song, F. & Gao, Y. (2012). Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern new jersey. *Atmospheric Pollution Research*, **3**, 247–257. 18

Rodrigues, A. & Diggle, P.J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics*, **37**, 553–567. 25

Rue, H. & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press. 51

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. 65

Ruf, T. (1999). The lomb-scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series. *Biological Rhythm Research*, **30**, 178–201. 11

Russo, A. & Soares, A.O. (2014). Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. *Mathematical Geosciences*, **46**, 75–93. 18

Ryu, D., Sinha, D., Mallick, B., Lipsitz, S.R. & Lipshultz, S.E. (2007). Longitudinal studies with outcome-dependent follow-up: Models and bayesian regression. *Journal of the American Statistical Association*, **102**, 952–961. 44, 73

Shi, J.P. & Harrison, R.M. (1997). Regression modelling of hourly NOx and NO2 concentrations in urban air in london. *Atmospheric Environment*, **31**, 4081–4094. 18, 22

Shin, H.H., Stieb, D.M., Jessiman, B., Goldberg, M.S., Brion, O., Brook, J., Ramsay, T. & Burnett, R.T. (2008). A temporal, multicity model to estimate the effects of short-term exposure to ambient air pollution on health. *Environmental health perspectives*, **116**, 1147. 17

Shumway, R.H. & Stoffer, D.S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics, Springer New York, 4th edn. 56

Song, X., Liu, Y., Hu, Y., Zhao, X., Tian, J., Ding, G. & Wang, S. (2016). Short-term exposure to air pollution and cardiac arrhythmia: a meta-analysis and systematic review. *International Journal of Environmental Research and Public Health*, **13**, 642. 18

Stoffer, D. (2017). *astsa: Applied Statistical Time Series Analysis*. 57

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 111–147. 27

Tómasson, H. (2015). Some computational aspects of gaussian carma modelling. *Statistics and Computing*, **25**, 375–387. 12, 43

Wang, Z. (2013). cts: An R package for continuous time autoregressive models via kalman filter. *Journal of Statistical Software*, **53**, 1–19. 53, 58

Weather Underground (2015). Site which provides weather data. 21

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bull. Internat. Statist. Inst.*, **39**, 105–129. 12