Spatio-temporal models for air pollution

submitted by

Uziel Gonzalez

to the

University of Exeter

as a thesis

for the degree of Doctor of Philosophy in Mathematics

May, 2022

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signature.....

Uziel Gonzalez

Contents

List of Figures

List of Tables

Summary

Air pollution is the biggest environmental risk to global health and it is estimated that, globally, 7 million deaths can be attributed to air pollution each year (WHO, 2018). The World Bank estimates that, in 2016, the overall cost of ambient air pollution to the global economy was an estimated US $5.7 trillion or 4.4 per cent of global GDP (World Bank, 2016). A number of different air pollutants have been associated with adverse health effects, including fine particulate matter ($PM_{2.5}$), nitrogen dioxide and ozone. In studies of the effects of air pollution, exposure information is often obtained from a fixed number of monitoring sites within the region of interest. However, an increasing number of models of air pollution are being used that provide estimates of concentrations. These are used to represent exposures at every location in a health study area, rather than just at a number of fixed measurement locations. Another use of modelling of air pollution is to provide short-term forecasts that can be used to inform the behaviour of vulnerable people.

In this thesis, we develop statistical approaches to modelling, and forecasting, daily concentrations of $PM_{2.5}$ in urban areas. We consider two different approaches, both in terms of model formulation and performing inference. The first approach is Dynamic Space-Time Models (DSTM). Under this framework, a *data* model relates observations (measurements) to a *process* model that specifies the dynamic evolution of the "true" underlying process. This approach is implemented using two different methods for estimation: methods of moments and expectation-maximisation. We also develop an approach using Bayesian Hierarchical Spatio-Temporal modelling (BHSTM). The inference is done using computational efficient methods for Bayesian inference (integrated nested Laplace approximations). This model allows predictions of daily $PM_{2.5}$ over both space and time, which can be used to interpolate both past measurements and future predictions. Both approaches were implemented using data from Greater London, with their performance evaluated in terms of their ability to predict daily concentrations of $PM_{2.5}$ over time at different measuring sites. Both methods were able to accurately predict future values of daily $PM_{2.5}$ at different locations, with one-day ahead predictions being more accurate than those used for longer periods, as might be expected. One of the major advantages of the BHSTM approach is that it provides a straightforward method for producing estimates of the uncertainty that is associated with predictions.

Acknowledgement

First and foremost, I have to thank my supervisor, Prof. Gavin Shaddick. Without his assistance and involvement in the process, this thesis would have never been accomplished. I would like to thank you for your support and understanding over these past years. I am also grateful for the financial support provided by CONACyT, Mexico.

I am immensely grateful to Dr. Sergio Juarez, who always encourages me to take these steps. Getting through my dissertation required more than academic support, and I have many people to thank for listening to me over the past years. I cannot begin to express my gratitude and appreciation for their friendship. I would like to thank my friend Cristina who opened both her home and heart to me when I needed it.

Most importantly, none of this could have happened without my family. My parents, who offered their constant encouragement and support through phone calls, always believe in me more than I do. With his own brand of humour, Miguel has been kind and supportive to me over the last several years. I also would like to thank Jorge for his advice and help in difficult times.

Elaine, every time I was ready to quit, you did not let me and I am forever grateful. This dissertation stands as a testament to your unconditional love, encouragement and patience.

Chapter 1

Introduction

Air pollution is a significant risk to public health and globally is considered to be the leading environmental risk to health Davies (2017); World Health Organization (2016); Royale College of Physicians (2016); Parliament UK (2018) The WHO has estimated that 4.2 million deaths per year worldwide are attributed to ambient (outdoor) fine particulate matter air pollution ($PM_{2.5}$). A problem of local and national concern is the large scale pollution of the atmosphere. Modern research into air pollution started from the middle of the twentieth century. Since then, there has been an increasing interest about the possible effects of air pollution on health. This interest was generated by several early air pollution episodes, one of them in Belgium (1930). In this case, 63 people died due to high concentrations of sulfur dioxide in the Meuse River Valley (Heimann, 1961; Ayres *et al.*, 1972; Pope *et al.*, 1995; Anderson, 2009). Another incident occurred in 1948 when 20 people died from cardiac and respiratory disease and approximately 6,000 people suffered from respiratory problems in Donora, Pennsylvania (Anderson, 1967; Snyder, 1994; Chew *et al.*, 1999). One final example occurred in 1950, when 22 people died and 320 were hospitalised mainly because of an accidental release of hydrogen sulfide at Poza Rica, Mexico (Collins & Lewis, 2000).

A number of different air pollutants have been associated with adverse health effects, including fine particulate matter, nitrogen dioxide and ozone. Particle matter (PM) consists of fine solid and liquid droplets, other than pure water, that are dispersed in air. PM originates from natural as well as anthropogenic sources. Natural sources include: wind-blown soil dust, volcanic ash, forest fires, sea salt and pollen. Anthropogenic sources include: thermal power plants, industries, commercial and residential facilities and motor vehicles which use fossil fuels. A major source of these particles, and the

other pollutants, are combustion processes, in particular diesel combustion. Many studies have implicated particle matter air pollution as contributing to the incidence and severity of respiratory disease Pope & Dockery (1992); Pope *et al.* (1995). Decreases in pulmonary function associated with particle pollution levels have been observed. PM pollution is a concern because it contains toxic metals and sulphuric acid, and it can penetrate the respiratory system. Several epidemiological studies have consistently reported associations between a variety of pollutants and both mortality and morbidity, Laden *et al.* (2000); Dominici *et al.* (2000, 2002).

Nitrogen dioxide is a primary gas released into the atmosphere after burning of fuel. It is a strong oxidising agent that plays an important role in atmospheric transformation reactions and is converted into gaseous nitric acid, toxic organic nitrates and tropospheric ozone (a major component of smog). It is identified as one of the important air pollutants having significant impacts on human health. High NO_2 concentration causes short-term exposure health problems, such as cardiovascular and respiratory diseases. Cancer and stunted mental development in children are examples of long-term exposure diseases related to this pollutant. There are natural sources of NO_2, such as atmospheric oxidation of ammonia, microbial activities in soil and lightning, but anthropogenic sources are more prevalent than natural sources. Due to the oxidation processes in the atmosphere, oxides of nitrogen and primary NO convert into NO_2. In the presence of O_3, the reaction happens at a high rate. In most of the cases, formation of NO_2 is under the influence of ground-level O_3.

Pollutants in general have varying atmospheric lifetimes, but all can persist in the environment, including PM. The atmospheric lifetime of particulate matter is strongly related to particle size, but may be as long as 10 days for particles of about 1mm in diameter Onursal & Gautan (1997); Richter & Williams (1998).

In Europe, there is a long history of national and international regulatory approaches to air quality management. Starting with the Clean Air Act of 1956 in the United Kingdom, successive legislation, including the European emission standards passed in the European Union in 1992, has succeeded in reducing ambient air pollution levels over time (Shaddick & Zidek, 2014; Turnock *et al.*, 2016; Guerreiro *et al.*, 2014; Kuklinska *et al.*, 2015).

In recognition of these risks associated with fine particle matter, national and international ambient air quality standards have been set for a range of pollutants. These standards include particle matter with an aerodynamic diameter equal or less than a $10\mu g/m^3$ and for even smaller particles, $2.5\mu g/m^3$, (PM_{10} and $PM_{2.5}$ respectively) as

well as other pollutants, including NO_2 and ozone.

Current statutory limits state that annual average concentrations of NO_2 and $PM_{2.5}$ should not exceed 40 $\mu g/m^3$ and 25 $\mu g/m^3$ respectively (European Commission, 2017). The World Health Organisation (WHO) Air Quality Guidelines (AQG), that are designed to protect the public from the adverse health effects of air pollution were, until recently, the same as the EU statutory limits for NO_2, stating annual averages should not exceed 40 $\mu g/m^3$, but much lower than the EU statutory limits for $PM_{2.5}$, stating annual averages should not exceed 10 $\mu g/m^3$ (World Health Organization, 2005). In 2021, the WHO re-issued the air quality guidelines and notably reduced guideline for annual average $PM_{2.5}$ from 10 $\mu g/m^3$ to 5 $\mu g/m^3$ (Weltgesundheitsorganisation *et al.*, 2021), reflecting the increasing evidence of adverse health effects of air pollution at lower levels.

Many epidemiological research studies across the world have shown that exposure to poor air quality impacts on people's health. Some effects are nearly immediate, known as acute effects, while some happen over a longer term, known as chronic effects.

The short terms effects of air pollution on health are especially great during air pollution "episodes" (Anderson, 1999). Periods of higher levels tend to occur in episodes that can last several days. These episodes often happen when the weather conditions cause pollutant concentrations to build up above normal levels. They can also be caused by the route the air has travelled over previous days. For example, if the air passes over Europe, it can pick up pollutants on the way. This is different from changes in local sources, i.e traffic on the road, which tend to remain fairly constant in general. In the past, episodes in the winter have been known as "winter smog" or "pea soupers". The most infamous incident caused by coal burning was the London smog in 1952. It is often referred to as the "Great Smog", in which levels of black smoke, an early measure of particulate matter, exceeded 4,500 $\mu g/m^3$. That episode lasted five days and has been associated with between 4,000 to 12,000 deaths (Polivka, 2018). These extreme episodes happened by burning huge quantities of domestic and industrial coal. However, due to the availability of mains natural gas and the Clean Air Acts, these kind of regular episodes seen in the 1950s have been eliminated, at least in the UK.

Nowadays, in London, air pollution is particularly high between 3 and 8 times a year. When this happens, air quality advice is broadcast by the Mayor and Transport for London across the city. This happens at tube stations, bus stops and along major roads. The aim is to allow the public, and particularly vulnerable people, to make informed choices. People decide how much time they spend outdoors, where they go,

and what they do. Daily information on air pollution in London can be found at www.londonair.org.uk, which also has information on predicted air pollution episodes.

The World Meteorological Organization (WMO) Global Atmosphere Watch (GAW) Programme runs the Global Air Quality Forecasting and Information System (GAFIS). It provides air quality forecasting and information systems, from the global level to more local levels for urban areas. Daily air quality forecasts at the global scale are provided by institutions such as the Copernicus Atmosphere Monitoring Service. They provide global air quality forecasts of the effects of anthropogenic emissions, desert dust and wild fires on air quality at a resolution of about 40 km (Zhang *et al.*, 2020). Regional forecasts can be more accurate than the global forecast, as they use more detailed spatial information of the emission, the chemistry and the transport. In North America and Europe, such regional models are already operational, including the DEFRA forecasts. In the UK, these forecasts are provided by the Met Office (https://uk-air.defra.gov.uk/forecasting). The Met Office model is a regional model that represents the background and regional air quality and has a 12km grid resolution. However, it does not represent the localised increases in pollution that happen in cities, roads or in urban centres.

Numerical models are based on existing knowledge of a number of factors that will affect air quality, including:

- Emissions of pollutants
- Transport and dispersion of pollutants by winds
- Chemical reactions amongst reactive gases and aerosols
- Removal processes, such as rain and deposition on surfaces.

These chemical transport models (CTMs) (Brauer *et al.*, 2003) predict concentrations of air pollutants over various spatial locations, but have a number of limitations. The numerical computer models that represent these deterministic models can take a long time to run and the uncertainty associated with these estimates is not generally available. Also, they generate outputs for grid cells which may be large (Kalnay, 2003) and although their output may be perfectly valid at that scale they do not reflect local spatio–temporal patterns in air pollution levels.

A different approach is to use statistical models that are based on local measurements of data and model the spatial and temporal relationships in data to produce predictions. These predictions can either be for previous time points, i.e. producing estimates of

underlying levels of pollution accounting for potential errors or anomalies in recorded data, or for filling in missing data, or for predicting future periods in time (forecasting). Unlike numerical models, statistical models are designed to provide estimates of uncertainty. Statistical models, in particular, spatio-temporal statistical models, are the focus of this thesis.

A traditional approach to performing spatial interpolation of air pollution data, in order to produce spatial predictions at unmeasured locations, has been kriging. For further details, see chapter 6. Some examples can be found in Cressie (1993), Chiles & Delfiner (1999), Stein (1999) Janssena *et al.* (2008) Montero *et al.* (2015). Bayesian versions of kriging have been developed and applied to air pollution data in a number of studies, for early examples, see Le & Zidek (1992); Handcock & Stein (1993); Ecker & Gelfand (1997). The use of Bayesian modelling to perform spatial modelling, based upon similar concepts to kriging includes Shaddick *et al.* (2013) who modelled NO_2 in Europe using MCMC (see section 3.4) and Vicedo-Cabrera *et al.* (2013) who modelled NO_2 and SO_2 in Italy to provide exposure estimates for a health study. Both of these last two examples used WinBUGS (Lunn *et al.*, 2000) to perform the MCMC.

The idea of extending spatial modelling to spatio-temporal modelling of air pollution within a Bayesian context has been explored by a number of authors. Furthermore, it has seen many advances in recent years. For example, Zidek *et al.* (2002) implemented an approach suggested by Le *et al.* (1997), to model the space-time field of daily PM_{10} in Vancouver, Canada. For simplicity, they analysed each monitoring site separately and chose an AR(1) model to represent the temporal structure. In this study, a hierarchical Bayesian model is proposed and the spatial covariance is not specified in the first level. Then, uncertainty about the spatial covariance is incorporated into a second level prior. In Shaddick & Wakefield (2002), a hierarchical dynamic linear model is proposed, with spatial structure on the residuals, to model daily observations of four pollutants at eight monitoring sites in London. They considered a separable model, implemented via Markov chain Monte Carlo sampling (using WinBUGS). Sahu *et al.* (2007) developed a spatio-temporal model for the analysis of daily ozone observations in Ohio, U.S. They fitted a stationary auto-regressive model for the temporal correlation and an exponential covariance structure for the spatial correlation. The model is specified within a Bayesian framework and is fitted using MCMC methods. In another example, (Lee & Shaddick, 2010) used a spatio-temporal model for four pollutants (CO, NO_2, O_3 and PM_{10}) in London, again using MCMC, incorporating the predictions into the model for analyses of health risks.

More recently, there has been great interest in using computational efficient methods

of performing Bayesian inference, notably INLA (Rue *et al.*, 2009) when fitting spatio-temporal models to air pollution data. Cameletti *et al.* (2011, 2013) used a separable spatio-temporal model to analyse PM_{10} data in the Po valley (northern Italy). They proposed a representation of a Gaussian Field with Matérn covariance function as a Gaussian Markov Random Field. This was done through the Stochastic Partial Differential Equations (SPDE) approach (Lindgren *et al.*, 2011). This is the approach used in chapter 7 and more details on the use of INLA and the SPDE approach can be found in Section 7.2.

In other examples, (Shaddick *et al.*, 2018b,a) used INLA when fitting a Bayesian hierarchical model to global air pollution that had nested random effects representing countries and region. They used the model to produce high-resolution estimates of $PM_{2.5}$ on a grid ($10^o \times 10^o$ resolution) that were used for health burden calculations. This model was extended in Shaddick *et al.* (2020b) to a spatio-temporal model, following the same structure as Cameletti *et al.* (2013) including a continuous spatial process and a random walk. This allowed for within grid-cell variability, where there was enough local data and the use of multi-year measurements. This model was used in Shaddick *et al.* (2020a) to estimate trends in global air pollution and to provide one of the United Nations Sustainable Development Goal (SDG) indicators related to air pollution: SDG 11.6.2 (annual average exposure to fine particulate matter air pollution in cities). Another example of using INLA in this context can be found in Morrison *et al.* (2016) who modelled the relationships between healthcare utilisation and air pollution in British Columbia, Canada. They used a multivariate temporal model to forecast dispensations of medication and visits to primary care (for asthma-related complaints) during increased levels of air pollution ($PM_{2.5}$) during a wildfire season.

1.1 Overview of thesis

The overall aim of this thesis is to develop and compare different prediction models for daily concentrations of $PM_{2.5}$. Two different approaches are explored to produce predictions of air pollution levels, using data from London to show how they can be implemented.

The two approaches that we consider are:

1. Dynamic Space-Time Models (DSTM)
2. Bayesian Hierarchical Spatio-Temporal Models (BHSTM)

The fundamental difference in these two approaches is how spatial dependence is in-

corporated into the models. In the first approach, dependencies between data from different locations is integrated into the transition matrix that governs how the underlying spatio-temporal process (the underlying pollution field) transitions from one time to the next. The DSTM approach has a number of advantages, including the fact that dependencies are based solely on correlations between the data and so can readily be used with large datasets that include both spatial data and complex interactions between multiple pollutants.

One disadvantage is the lack of a defined spatial structure, i.e. a model that determines the spatial relationships between data from different locations. This means that it is not straightforward to perform spatial prediction, i.e. predict levels of pollutants at unmeasured locations, which would enable maps to be produced. This ability is very useful in understanding spatial patterns in pollutants and in being able to assign levels of pollution to subjects in epidemiological studies.

In the second approach (BHSTM), spatial dependence is explicitly modelled, which enables such prediction (at unmeasured locations) but does require knowledge of the nature of the dependence, which, if not known entirely (which will usually be the case), will mean making assumptions that simplify the process. Pollutants such as $PM_{2.5}$ are complex combinations of particles and therefore, making such assumptions can be restrictive and possibly unrealistic. For example, assuming space-time anisotropy into a model to analyse $PM_{2.5}$ is unrealistic and often there may not be enough data to fully justify all assumptions. However, in all modelling it is necessary to take a pragmatic approach: balancing the need to make such assumptions, in order to achieve the goals of the analysis and produce useful results, e.g. spatial mapping, while respecting the complexity of the underlying processes being studied (in this case air pollutants) and the data it generates.

In this thesis, we investigate the use of these two different approaches in the temporal and spatial prediction of air pollutants and highlight the advantages, and disadvantages, of each. In each case, we evaluate the ability of the different methods to produce predictions and short-term forecasts of $PM_{2.5}$. Depending on the method being considered, we perform evaluation based on the models ability to predict over time and/or space. For example, in temporal prediction, three days of data from a specified time period are omitted from the dataset used to fit the model, and the model is then used to predict the concentrations on those days. The accuracy of the predictions is assessed by calculating the *mean squared prediction error* (MSPE).

Throughout the thesis, these models are implemented, using data from the London

Air Quality Network. Then, they are compared with respect to their ability to model spatio-temporal patterns in the data. Following this introduction, chapter 2 provides some basic exploratory analysis on this to gain insights and to identify some trends and outliers in the air pollution data from London.

The remainder of the thesis is organised as follows. Chapter 3 introduces the concepts of hierarchical modelling, including both frequentist and Bayesian approaches. This chapter also includes a background to Bayesian inference, which is the basis of the latter chapters of the thesis. In chapter 4, an overall view of spatio-temporal statistical modelling is given, including details of covariances, semi-variograms and an application of empirical based approaches of these to the London data.

In chapter 5, the framework for dynamical spatio temporal models (DSTM) is presented. DSTMs are different from the approaches described in chapter 6 and 7, which are largely based upon the characterisation of the covariance and how it represents dependencies structure in space and time. Some methods for inference, such as Bayesian inference using INLA (see chapter 7 and (Rue *et al.*, 2009) for more details) can be computationally efficient. However, in some cases it may be more efficient to use dynamical models to characterise (spatio-temporal) dependencies, leading to dynamic spatio-temporal models (DSTMs) (Wikle & Hooten, 2010).

One of the big challenges with the specification of such models has been the curse of dimensionality. Even with fairly simple models (for example, linear models based on first-order Markovian temporal structure with Gaussian errors), they can be over parameterised. One solution is to use a hierarchical model which helps with this by allowing dependency among groups of parameters. Within a DSTM, the main objective is to make inference on the unobserved states, which are the underlying process from which measurements are made (with noise). The first step is choosing an adequate DSTM representation, then the estimation is done by computing the conditional distributions of the quantities of interest, given the available information and using the Kalman filter or smoother. In chapter 5, details of estimation for linear DSTMs with Gaussian errors are given based upon methods developed for state-space models with consideration of methods for dimension reduction. A selection of methods for inference (including methods of moments and an EM algorithm) are implemented using the London air quality data and evaluated in respect to their abilities to produce accurate short-terms forecasts of $PM_{2.5}$. An advantage of this approach is the ability to analyses multiple pollutants, for example $PM_{2.5}$ and PM_{10}, NO_2) simultaneously in a computationally efficient manner, without the need to specify the possibly complex relationships between them (and over time and space).

In chapter 6, covariance-based methods for spatio-temporal modelling are described, including *kriging*. It is one of the most commonly used methods used for spatial prediction, and has been extended for the spatio-temporal case. Kriging is a covariance-based method that can be used to interpolate, in which interpolated values are modelled by a Gaussian process. Different types of kriging models are discussed, including simple kriging, ordinary kriging and universal kriging, and their spatio-temporal extension. This chapter also contains a discussion of kriging methods that incorporate dimension reduction, including fixed rank kriging (FRK). The chapter includes the application of spatio-temporal kriging and FRK to the London air quality data. They have shown their ability to produce maps of air pollution based on measurements at a distinct number of locations.

In chapter 7, we describe a spatio-temporal model within a Bayesian hierarchical modelling framework. It is based upon a Gaussian spatial field, in which spatial structure defined by its covariance and a state process is characterised by a first order autoregressive process. To allow for efficient computation when doing Bayesian inference, we use the stochastic partial differential equation (SPDE) introduced by Lindgren *et al.* (2011). This model is applied to the $\log(PM_{2.5})$ data from London and evaluated with respect to its ability to predict accurately over time and space. This approach requires the spatial relationship between data from different locations to be explicitly stated, i.e. the form of the relationship between correlation and distance (between locations), which will require making certain assumptions as described above) but does mean that spatial prediction is possible and thus it can be used to produce maps of pollution over the study area.

Finally, chapter 8 provides a summary of the key findings from this thesis, together with overall conclusions and suggestions for future research.

Chapter 2

Air pollution monitoring in London

The amount of monitored pollution data that is available is increasing around the world. In London, air pollution remains a serious problem. Workplaces and people are concentrated not only in the heart of the city, but also in surrounding areas. As one of the largest cities in Europe, its data is provided by the London Air Quality Network (LAQN). This was formed in 1993 to coordinate and improve air pollution monitoring in London. The majority of London's 33 boroughs supply measurements to the network with additional measurements from local authorities surrounding London, thereby providing an overall perspective of air pollution in London and the Home Counties. The LAQN is operated and managed by the Environmental Research Group (ERG) at Imperial College London.

The **R** package ggmap together with Open Street or Google Maps can be used to display spatial data. London suffers from high levels of pollution in a similar way to most major UK cities. Due to the interest in reducing these levels, it is natural that London and its surroundings have an extended air quality network.

NO_2, PM_{10} and $PM_{2.5}$ daily observations will be used in 19 different stations across the Greater London area from 01 August 2015 to 15 March 2016. All data is accessible trough the **R** package openair, (R Core Team, 2020; Carslaw, 2019; openair web page, 2021). This package was developed for the purpose of analysing atmospheric composition data, Carslaw & Ropkins (2012). It includes some features such as access to UK air pollution monitoring sites and utility functions.

The **R** package ggplot2 can also be used to display spatial data. London is currently the main area failing to comply with the legally binding limits set by the EU, London Air Quality Network (2018). The size of the city, along with a dense road network and high buildings, means that central London tends to be one of the most polluted cities in the UK. Pollution can build up in London when it becomes trapped between buildings, or in the local area during invariable weather. Most of the monitoring sites are located in the central area along the River Thames. The 19 sites part of the air quality network in Greater London can be seen in figure 2-1 and their respective coordinates, code names and site type in table 2.1.

Some considerations have been taken for the selection of the monitoring sites. The first condition is that sites must lie between latitude: 51.0 to 52.0 and longitude: -0.6 to 0.4. Another condition is that $PM_{2.5}$ must be measured at the monitoring location. Although there are a few more sites monitoring this pollutant, these ones does not have measurements during the period of interest, therefore they were discarded. The period of interest includes from 01/August/2015 to 14/March/2016 and three more days to evaluate the forecasts. Two more intervals of time were considered, from 12/August/2015 to 25/March/2016 and from 01/March/2015 to 13/October/2015. Finally, it should be notice that all the observations were imported from openair through the importKCL function.

2.1 Exploratory Analysis

It is a common practice in meteorology to consider log-scale for pollutant observations, in particular PM data is measured in $\mu g/m^3$ units. The reason for making a logarithmic transformation of the raw data and working with the log of PM is that PM daily measurements do not follow a Gaussian distribution. This convention will be followed in this work for fitting and plotting purposes, all comparisons were made in the original scale.

The locations of the monitoring sites to be analysed can be seen in Figure 2-1 and their respective coordinates in table 2.1. As one can observe, most of these sites are located in the central area of London, the coordinates are also shown. The codes and names correspond to those sites considered for this thesis, it is accessible through openair. The administrative boundaries for London boroughs in figure 2-1 were taken from the Greater London Authority (GLA). The London data tore is an open data-sharing portal where free data relating to the capital can be found.

In trying to explore the data, there are particular challenges in spatio-temporal data vi-

Longitude	Latitude	Code	Site name	Site type
-0.125848	51.522287	BL0	Camden-Bloomsbury	Urban Background
0.184877	51.465983	BX1	Bexley-Slade Green	Suburban
-0.175284	51.544219	CD1	Camden-Swiss Cottage	Kerbside
-0.128774	51.527975	CD9	Camden-Euston Road	Roadside
-0.077765	51.513847	CT3	City of London-The Aldgate School	Urban Background
0.074003	51.490532	GN0	Greenwich-A206 Burrage Grove	Roadside
0.017697	51.492571	GN2	Greenwich-Millennium Village	Industrial
0.095111	51.486957	GN3	Greenwich-Plumstead High Street	Roadside
0.070766	51.452580	GR4	Greenwich-Eltham	Suburban
0.040725	51.456357	GR9	Greenwich-Westhorne Avenue	Roadside
-0.068218	51.599302	HG1	Haringey-Haringey Town Hall	Roadside
-0.298775	51.617327	HR1	Harrow-Stanmore	Urban Background
0.205460	51.520787	HV1	Havering-Rainham	Roadside
-0.213492	51.521046	KC1	Kensington and Chelsea-North Ken	Urban Background
-0.178809	51.495503	KC2	Kensington and Chelsea-Cromwell Road	Roadside
-0.441627	51.488780	LH0	Hillingdon-Harlington	Urban Background
-0.039641	51.474954	LW2	Lewisham-New Cross	Roadside
-0.141661	51.389286	ST5	Sutton-Beddington Lane north	Industrial
-0.008418	51.515046	TH4	Tower Hamlets-Blackwall	Roadside

Table 2.1: Coordinates, codes names and site type of the air quality network in Greater London composed by 19 monitoring sites.

sualisation, since one has at least three dimensions to consider. Exploring data through visualisation or summaries can be difficult, since one has one or more dimensions in space and one in time.

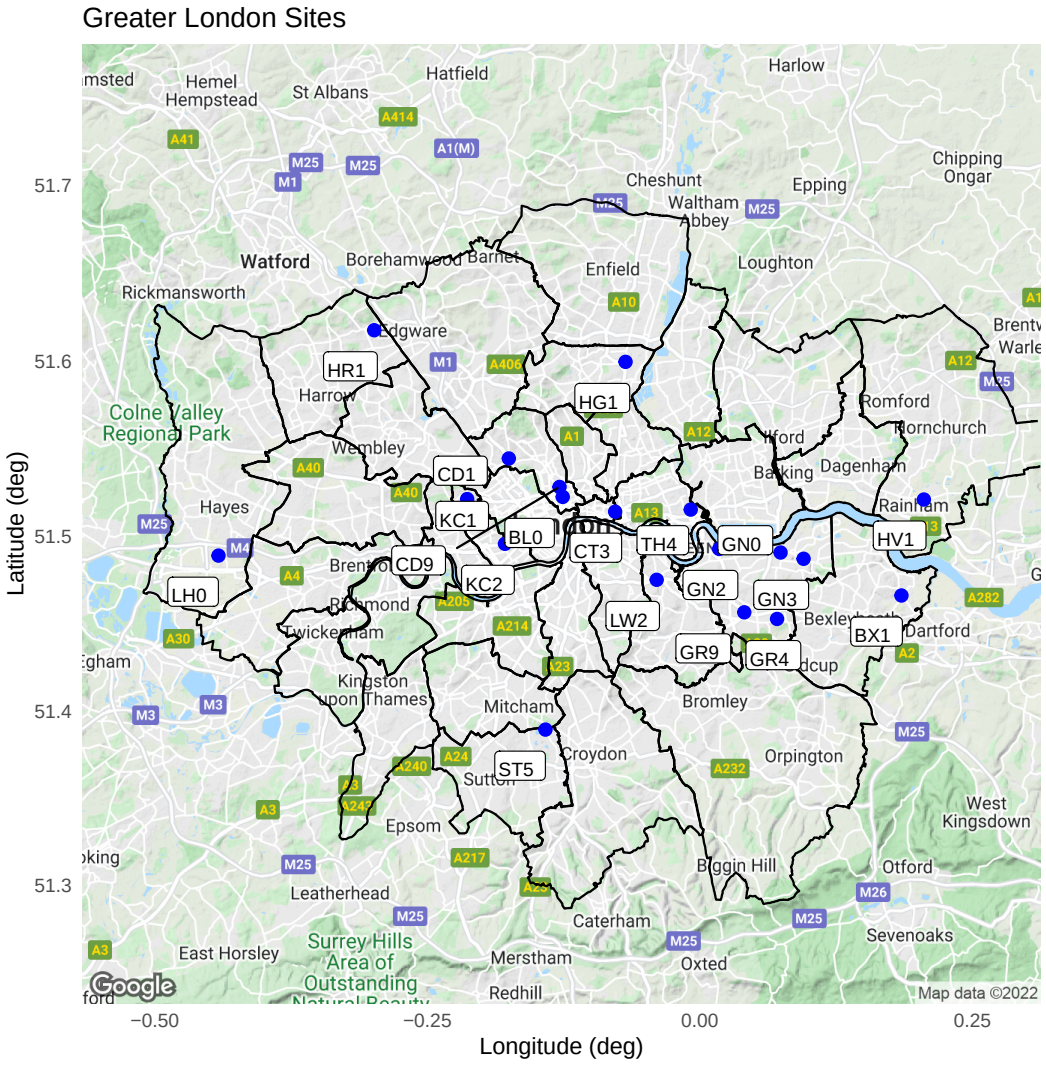

Figure 2-1: Air quality network in Greater London composed by 19 monitoring sites obtained from the LAQN.

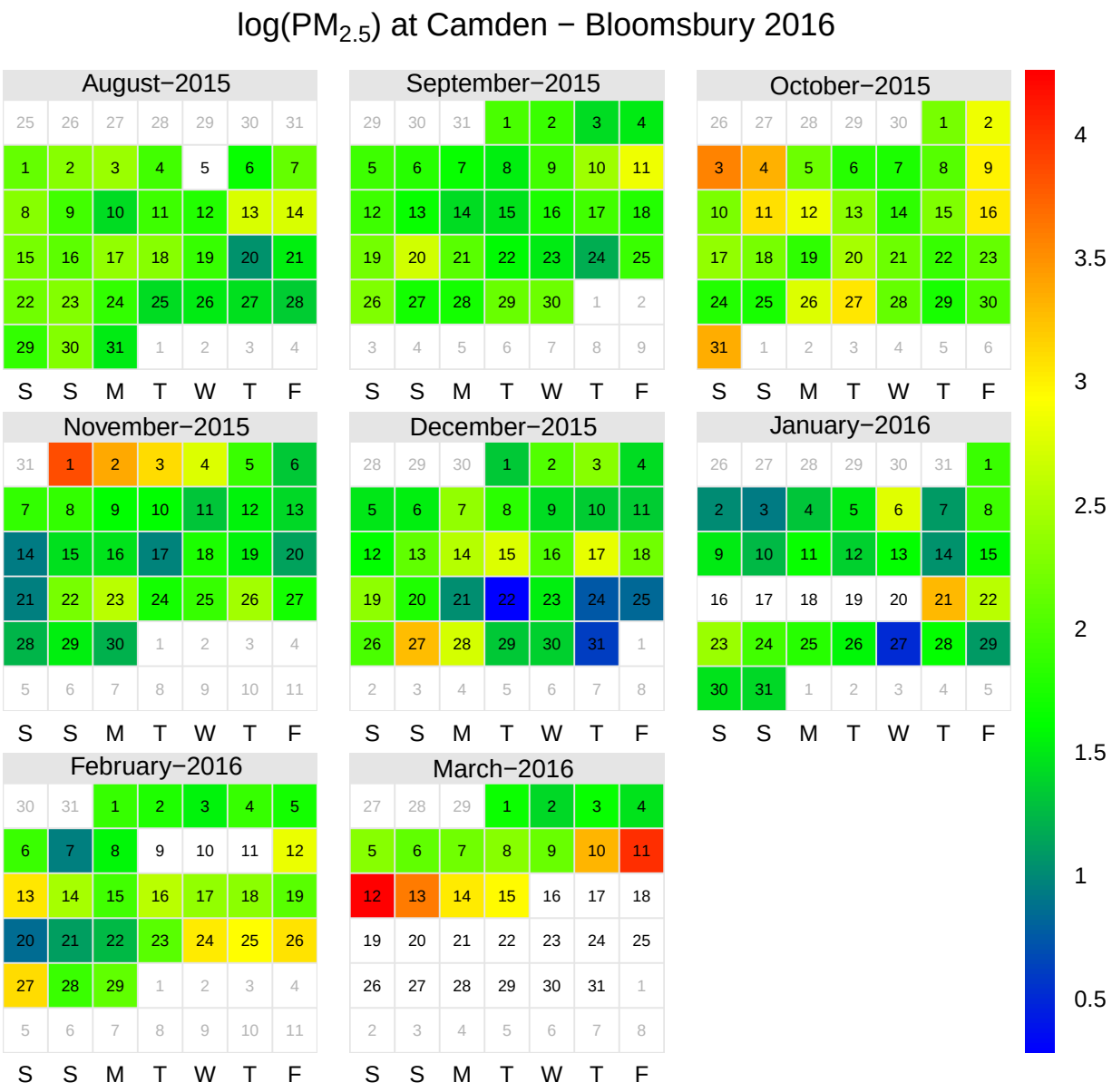

Figure 2-2: Calendar plot for log-scale $PM_{2.5}$ concentrations from 01/August/2015 to 15/March/2016 at Camden-Bloomsbury (BL0), one of the sites from the air quality network in Greater London area.

Sometimes, it is useful to visualise data in a variety of ways and this can be done through some functions available in openair. Daily concentrations can be shown in a calendar format, see figure 2-2. The idea is to provide some information on meteorological conditions for each day. This kind of plot can also help to highlight those conditions where daily maximum concentrations are above a particular threshold. In particular, in 2-2 one can observe that high concentrations in $PM_{2.5}$ correspond to some days during Autumn and Spring. One can also note that there were low level of $PM_{2.5}$ concentrations from August to September during 2015; a similar behaviour was observed for the rest of the sites.

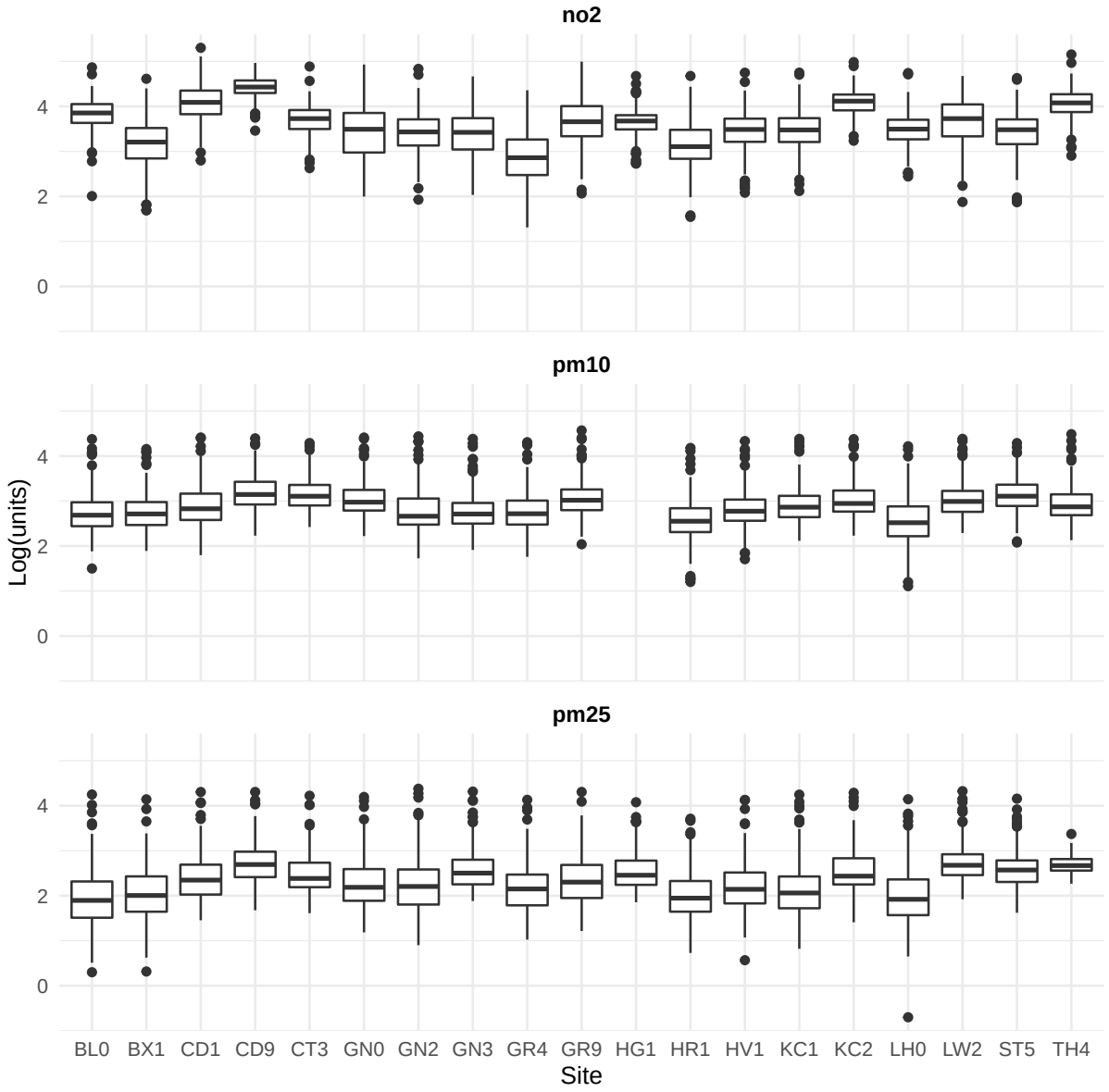

Figure 2-3: Boxplots: sites summaries across pollutants from 01 August 2015 to 15 March 2016 from the air quality network in Greater London area.

It could be helpful to generate boxplots across pollutants for each site, see Figure 2-3. Concentrations of NO_2, fine particle matter PM_{10} and $PM_{2.5}$ in log-scale are shown for 19 sites in Greater London area. The boxes represent the interquartile range from the $25th$ to $75th$ percentile and the empirical average is also shown. For the NO_2 levels, one can observe that concentrations are more widely dispersed across the sites, in contrast with the particle matter pollutants. Differences between PM_{10} and $PM_{2.5}$ are small across the sites; this can be explained due the high correlation. In particular, the monitoring site based on Camden (BL0), Greenwich-Millennium Village (GN2) and Harlington (LH0) show a wide range in concentrations for all these pollutants. In general, just a few outliers are observed for this data set based on the figure 2-3.

Visualisation is an important preliminary task that needs to be carried out before the data-analysis stage and the modelling stages. In this case, data is collected at stations that are fixed in space. According to the Figure, 2-4 one can get an idea of the overall spatio-temporal variation of the observed data. In the same figure, Greater London boundaries are plotted with data locations to give a better geographic perspective. One can see that the data set contains data for all pollutants (NO_2, PM_{10} and $PM_{2.5}$) during 07, 08, 09, 10, 11, 12, 13 and 14 March 2016.

Calculating trends for air pollutants is one of the most important and common tasks to be done in statistical analysis. Some trends can be observed from plotting the $PM_{2.5}$ concentration levels versus time, such analyses are useful for understanding how concentrations have changed through time. However, if one is interested in understanding why trends are as they are, more statistical analysis should be done. In Figures 2-5, 2-6 and 2-7 one can observe the time series associated with the log-scale of NO_2, PM_{10} and $PM_{2.5}$ in the data set respectively. They correspond to observations from 01 August 2015 to 14 March 2016 for all 19 locations. Due the time series are relatively short, visually it is not possible to detect any clear trend.

2.2 Interpolation methods

Spatio-temporal interpolation involves predicting the concentration at locations using a sample set of measured points from within the same area and period of time. One advantage of interpolation is that it allows us to create a continuous air pollution surface on the basis of relatively few point locations/time.

The estimation of concentrations at these locations is usually based on information from the nearest available measuring sites/times. Interpolation can be based on either deterministic or stochastic models. Deterministic methods considers mathematical functions

to interpolate a surface from the measured points, while stochastic methods incorporate randomness and apply statistical models to the measurements to create the surface.

In Burrough *et al.* (1998) two main types of deterministic methods are considered: local and global interpolators. Local interpolators calculate predictions on the basis of the nearest (in space or time) measurement points. In contrast, global interpolators fit a smooth surface defined by a function to the input data points across the whole study area. Global interpolators are mostly used for identifying the effects of global variations or trends in the data. Few studies have used deterministic methods to predict air pollution concentrations. For example, Abbey *et al.* (1993) used the *Inverse Distance Weighting* (IDW) to derive NO_2 concentrations. In Li *et al.* (2014), $PM_{2.5}$ data interpolation is conducted in the continuous space-time domain by integrating space and time simultaneously, using an extension approach. A main disadvantage of deterministic methods is that they do not provide estimates of uncertainty. In contrast, geostatistical methods can estimate both the predicted values and estimate the associated standard errors.

Inverse Distance Weighting (IDW)

An important statistical characteristic of spatio-temporal data is that nearby observations tend to be more alike than those far apart. One could find specific examples when the opposite can happen, but in general spatio-temporal data should be analysed jointly.

There are some cases where observations over time are unequally spaced or where there are missing observations. In these cases, an interpolation (in space and time) algorithm could be applied. For two-dimensional spatial data that are on an irregular lattice, smoothed plots can be created.

One of the simplest ways to perform spatio-temporal prediction would be averaging the data to give more weight to the nearest observations in space and time. One option to do this is through *inverse distance weighting* (IDW). If one has spatio-temporal data given by

$$\{Z(s_{11};t_1), Z(s_{21};t_1), \ldots, Z(s_{m_1 1};t_1), \ldots, Z(s_{1T};t_T), Z(s_{2T};t_T), \ldots, Z(s_{m_t T};t_T)\},$$

where for each time t_j one has m_j observations. The IDW predictor at some location

$\mathbf{s}_0$ and time t_0 is given by

$$\hat{Z}(\mathbf{s}_0; t_0) = \sum_{j=1}^{T} \sum_{i=1}^{m_j} w_{ij}(\mathbf{s}_0; t_0) Z(\mathbf{s}_{ij}; t_j), \tag{2.1}$$

where in this case $t_1 \leq t_0 \leq t_T$ and

$$w_{ij}(\mathbf{s}_0, t_0) \equiv \frac{\tilde{w}_{ij}(\mathbf{s}_0; t_0)}{\sum_{k=1}^{T} \sum_{l=1}^{m_k} \tilde{w}_{lk}(\mathbf{s}_0; t_0)}, \tag{2.2}$$

$$\tilde{w}_{ij}(\mathbf{s}_0; t_0) \equiv \frac{1}{d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))^{\alpha}}, \tag{2.3}$$

$d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))$ is the distance between the location $(\mathbf{s}_{ij}; t_j)$ and the prediction location $(\mathbf{s}_0; t_0)$, and α is the positive power coefficient. This parameter controls the smoothing and often $\alpha = 2$, often referred to as the bandwidth parameter. The IDW is a weighted average of the data points, giving the closest locations more weight. The distance $d(\cdot; \cdot)$ is usually the Euclidean, but another one can be considered, in particular if time and space need to be treated differently.

The IDW is a particular case of a spatio-temporal *kernel predictor*, that is

$$\tilde{w}_{ij}(\mathbf{s}_0; t_0) = k((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0); \theta), \tag{2.4}$$

where $k((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0); \theta)$ is a *kernel function* that quantifies the dissimilarity between $(\mathbf{s}_{ij}; t_j)$ and $(\mathbf{s}_0, t_0)$ and θ is a *bandwdith* parameter. An example of a kernel function is the *Gaussian radial basis kernel*,

$$k((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0); \theta) \equiv \exp\left(-\frac{1}{\theta} d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))^2\right) \tag{2.5}$$

where θ is proportional to the variance parameter in a Gaussian distribution.

One of the main criticism about IDW is that space and time have the same treatment, of course this is not a sensible assumption for most of the cases. A simple variation of this approach is proposed in Li *et al.* (2014). The idea is to adapt the traditional spatial IDW method and consider the spatio-temporal interpolation formula based on a extension of that approach. In this case, a different distance for 2.3 is considered,

$$d^*((\mathbf{s}_{ij}; ct_j), (\mathbf{s}_0; ct_0))^{\alpha},$$

where $c > 0$. The choice of the factor c is commonly empirical, based on observation

or experience. In particular, the proposed c factor is,

$$c = \frac{1}{T-1} \frac{(lon_{\max} - lon_{\min}) + (lat_{\max} - lat_{\min})}{2},$$

where lon and lat are the longitude an latitude respectively; and max and min denote the maximum and minimum respectively. When performing a spatio-temporal analysis using these methods, it is important that both space and time are on the same scale to avoid patterns in one domain dominating the other based solely on the scale of the data. Here, the scaling is controlled by the parameter, c, and the idea is to make the temporal scale equal to the average range of the spatial differences.

2.2.1 Implementation using the LAQN data

IDW is not just helpful for prediction, in practice it can also be used to generate a smoothed visual representations of certain data in a specific time. In other words, it can be used to obtain a smoothed spatial grid without time variance. In figure 2-8, one can observe a smoothed spatial plot of 15 of March 2016. The plot is generated with inverse distance power $\alpha = 2$, a Gaussian radial basis kernel with bandwidth $\theta = 0.5$ and considering the data from 15 March 2015.

As with any other prediction, IDW must be used carefully. The $15th$ panel in figure 2-9 shows the smoothed levels of $PM_{2.5}$ levels for 15 of March 2016 considering the data from 01 August 2015 to 15 March 2015 data set. This plot was obtained using IDW with $\alpha = 2$ and $\theta = 0.5$. If the prediction is compared with figure 2-8, it is clear that is over smoothed. Setting α to a greater value would give a less smooth surface since less weight is given to observations that are distant from the prediction locations. The predictions are similar considering different settings for θ.

2.3 Discussion

In this chapter, we have performed initial data analysis on data from the LAQN and specified the study period and spatial domain of the analyses that will be performed within the thesis. This resulted in the extraction of a subset of the available data, based on pollutant, time and space. This was followed by a number of ways of visualising the data, which is an important preliminary task that can give insights into the data and help to inform the more complex modelling that will follow. This understanding can also be gained by implementing simpler statistical approaches (than the DSTM and BHSTM that are the focus of the work in this thesis). One such approach that was used here to assess spatial patterns in the data is Inverse Distance Weighting (IDW),

which is based on estimating relationships between data points empirically. Although useful for visualising patterns, IDW has a number of disadvantages, in that correlations over both space and time are estimated in the same way, i.e. there is no allowance for possible differences in scale, which may not be a tenable modelling assumption; it does not allow the production of a forecast over time, nor for uncertainties to be estimated. However, it can be a useful option if one is interested in assessing spatial patterns, and in performing simple spatial prediction, as IDW is relatively easy to implement.

In the next chapter, we move from initial data analysis and visualisation with the implementation of empirical methods, to an introduction of more formal modelling, namely using Bayesian Hierarchical Models.

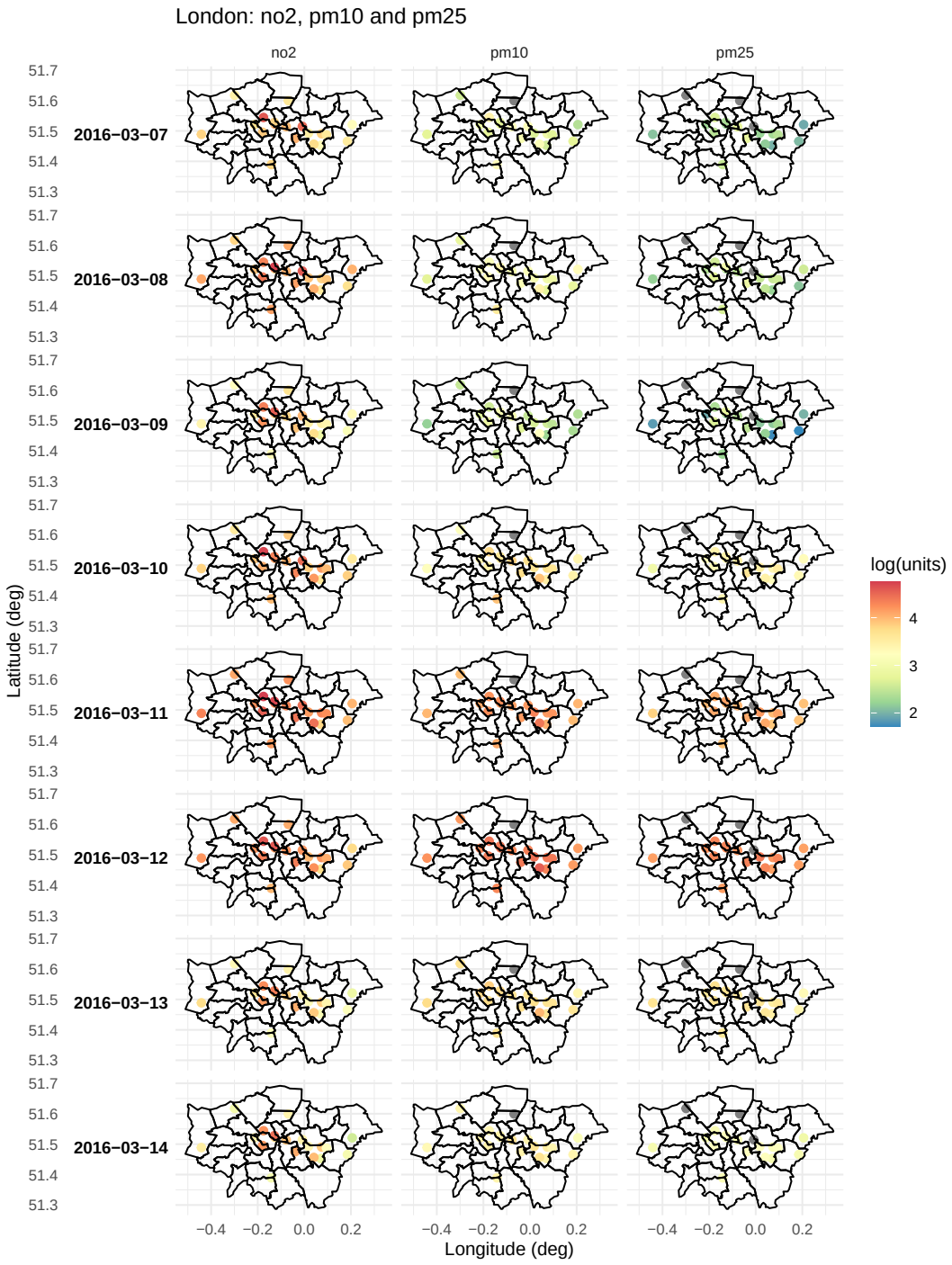

Figure 2-4: NO_2, PM_{10} and $PM_{2.5}$ levels in log-scale during 07, 08, 09, 10, 11, 12, 13 and 14 March 2016 from the air quality network in Greater London area. The points correspond to each monitoring site and the grey points are missing values.

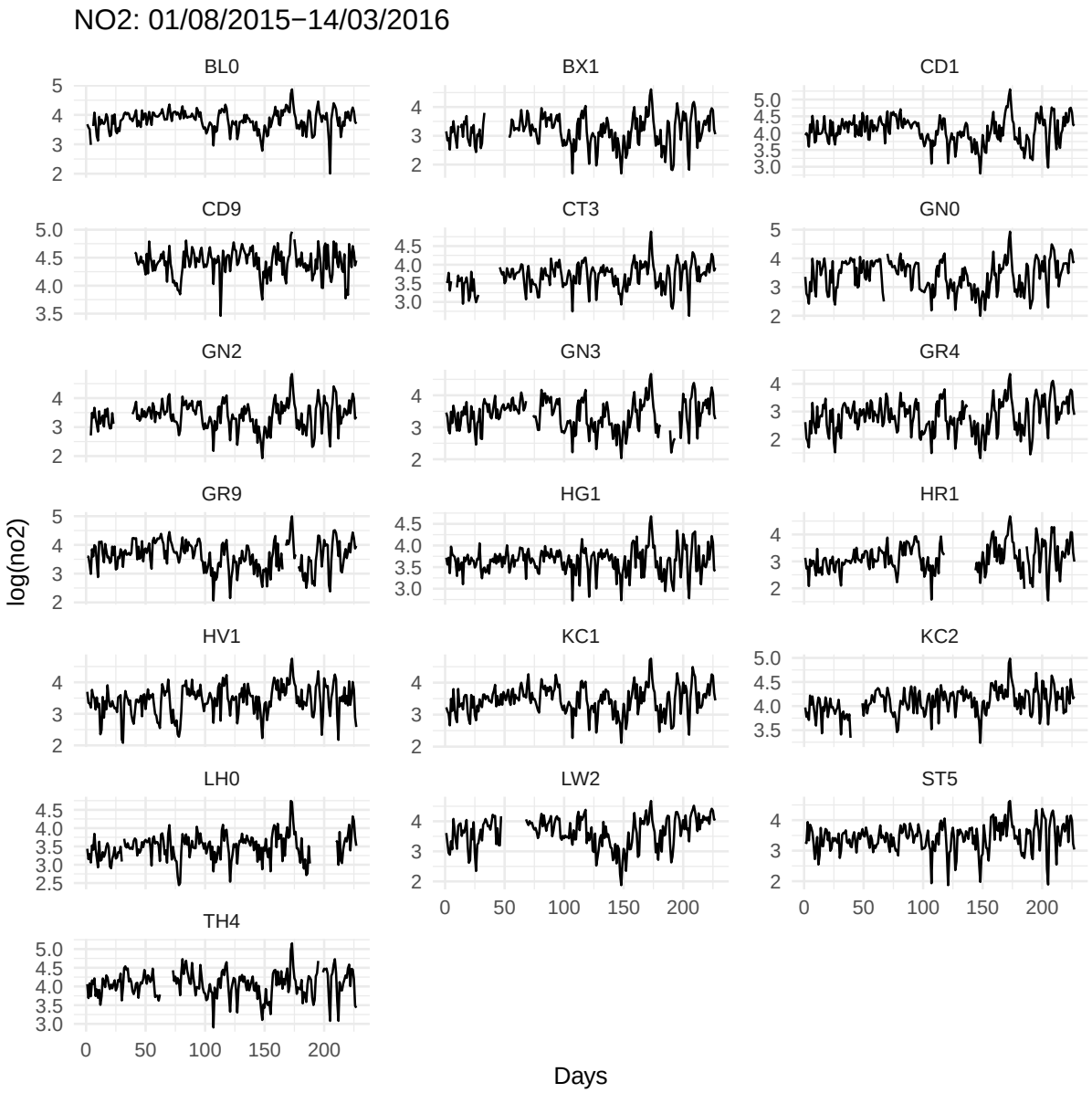

Figure 2-5: NO_2 time series from 01 August 2015 to 15 March 2016 for all sites from the air quality network in Greater London area.

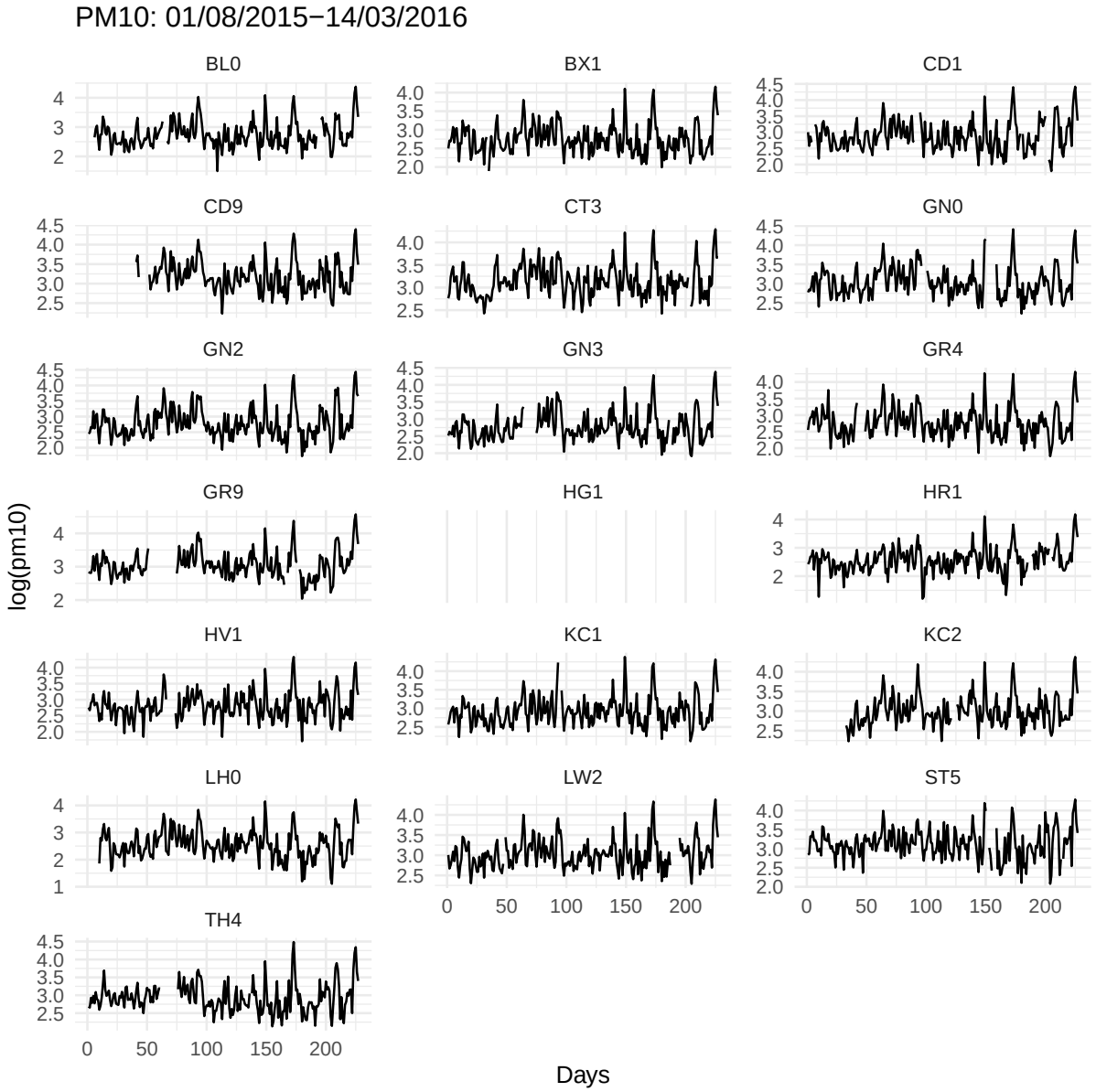

Figure 2-6: PM_{10} time series from 01 August 2015 to 15 March 2016 for all sites from the air quality network in Greater London area.

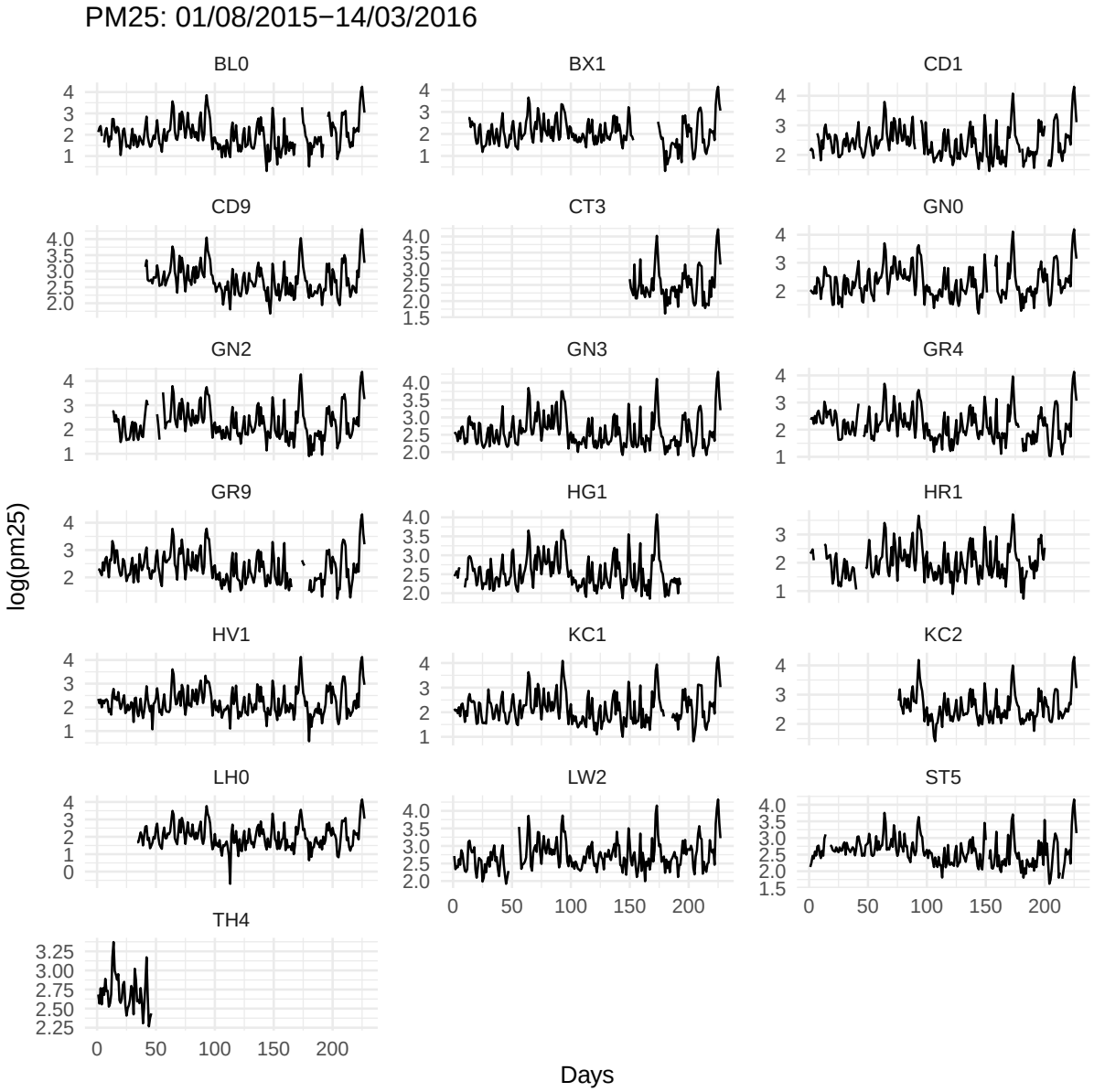

Figure 2-7: $PM_{2.5}$ time series from 01 August 2015 to 15 March 2016 for all sites from the air quality network in Greater London area.

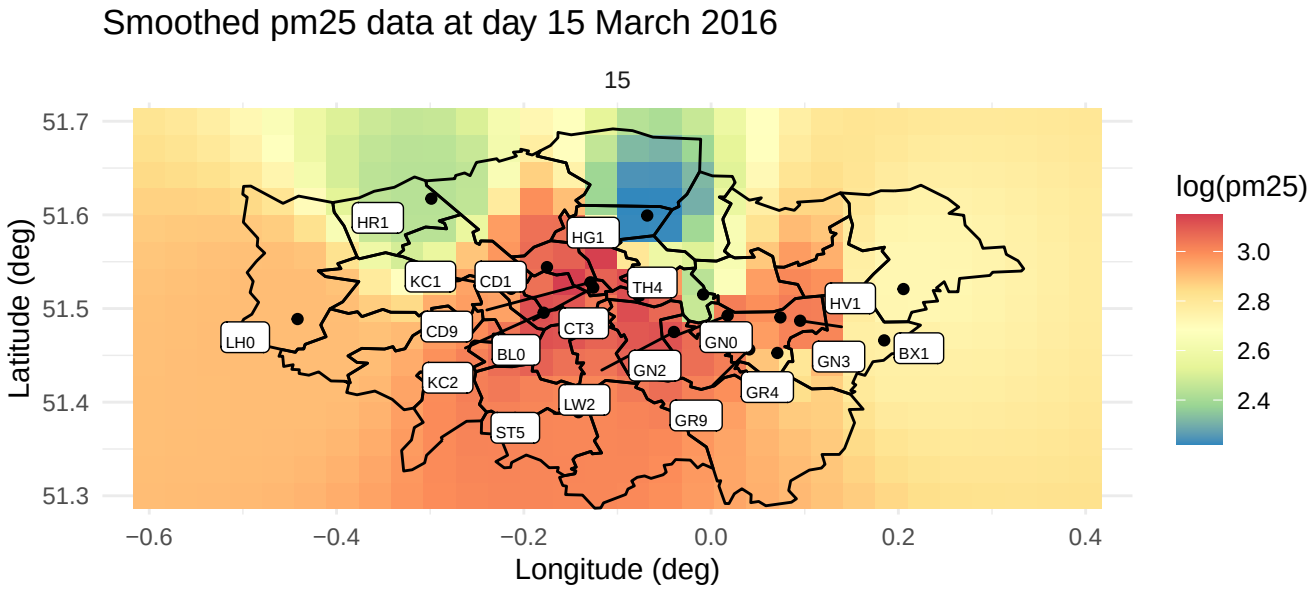

Figure 2-8: Grid with $PM_{2.5}$ levels in 15 of March 2016, Greater London. The panel was generated through the inverse distance weighting (IDW) with inverse distance power $\alpha = 2$, and a Gaussian radial basis kernel with bandwidth $\theta = 0.5$. The model was fitted using data from 15 March 2015.

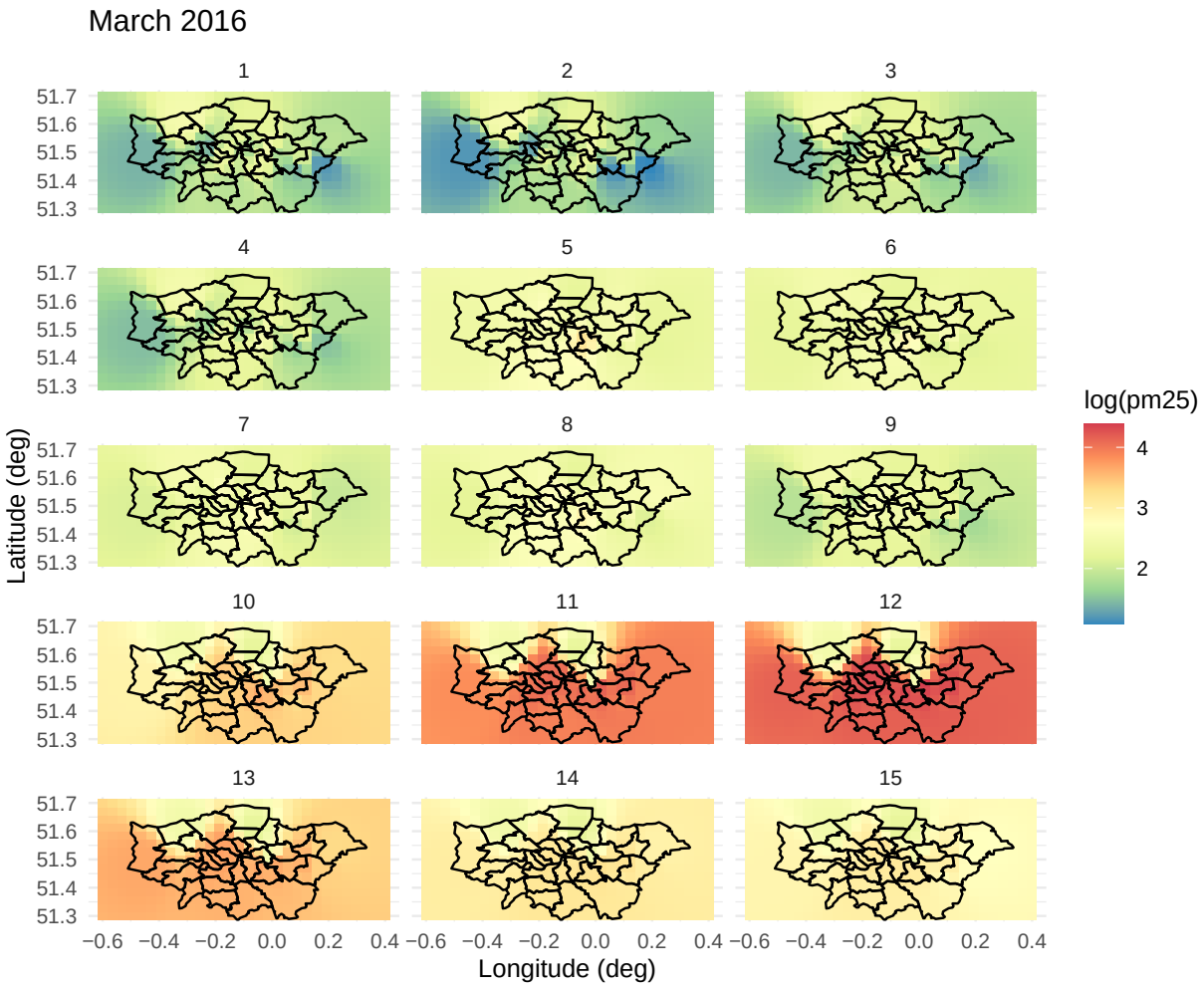

Figure 2-9: IDW smoothed levels of $PM_{2.5}$ levels, spanning the temporal interval of the data, 01 March 2016 to 15 March 2016. The time considered to fit the model is 01 August 2015 to 15 March 2016. It was generated with inverse distance power $\alpha = 2$, and a Gaussian radial basis kernel with bandwidth $\theta = 0.5$.

Chapter 3

Hierarchical modelling

In the following, we follow the terminology proposed by Berliner (1996), who defined a Hierarchical Model (HM) to include a data model, a process model, and a parameter model.

At the top level is the *data model* which expresses the distribution of the data given a latent process. Underneath the data model is the *process model*. Unknown parameters can be present, in both the statistical model for the data, conditioned on the process, and the statistical model for the process. When the parameters are given by prior distributions at the bottom level of the hierarchy, the HM is called Bayesian hierarchical model (BHM) (Gelman *et al.*, 2013; Shaddick & Zidek, 2015; Congdon, 2019). This probability model at the lowest level can be called *parameter level*.

Another advantage of the data-process-parameter modelling paradigm in an HM is that conditional-dependence structures usually come naturally. For example, it is often reasonable to assume that the data covariance matrix (given the corresponding values of the hidden process) is a diagonal matrix of measurement-error variances. This admits the process covariance matrix to capture the "pure" spatio-temporal dependence. Often this knowledge comes from physical or mechanistic knowledge.

The product of the conditional-probability components of the HM gives the joint probability model for all random quantities. The HM could be either a BHM or an EHM. It depends on whether a prior distribution is assigned on the parameters or the parameters are estimated respectively. A hybrid situation arises when some but not all parameters are estimated and the remaining have a prior distribution put on them.

The BHM also allows one to find the posterior distribution of the parameters given

the data, whereas the EHM requires an estimate of the parameters. Predictive and posterior distributions are obtained using Bayes' rule as it will be seen below.

If it is agreed the notation $[A]$ and $[A|B]$ for marginal and conditional probability of A given B, respectively, then the joint distribution of A and B can be written as

$$[A, B] = [A|B][B], \tag{3.1}$$

and the law of total probability can be written as

$$[A] = \int [A|B][B]dB, \tag{3.2}$$

where it is recalled that $\int g(B)[B]dB$. It denotes the expectation, or a summation in the case is where B is a discrete random quantity, of some function $g(B)$ of B. In terms of this notation, Bayes' theorem can be written as

$$[B|A] = \frac{[A|B][B]}{\int [A|B][B]dB} = \frac{[A|B][B]}{[A]}. \tag{3.3}$$

The predictive and posterior distributions cannot be calculated in closed form, in which case one relies on computational methods to deal with it. Nowadays, Monte Carlo samplers from a Markov chain whose stationary distribution the posterior distribution of interest have been used for this purpose (Gelfand & Smith, 1990). These Markov chain Monte Carlo (MCMC) methods have transformed the use of HMs for complex modelling applications, such as those found in spatio-temporal statistics.

3.1 Empirical hierarchical models (EHM)

The idea in this alternative approach is to assume that the parameters are fixed in the top two levels using the data. In this case, the model is known as an empirical hierarchical model (EHM). In certain cases the EHM can be preferred, one could choose to avoid informative priors on parameters. For further details see Cressie & Wikle (2011) and Jones (2001). This approach can be useful if there is lack of knowledge of the process or if computational efficiencies are favoured.

The following model is called a EHM:

An EHM uses just the first two levels, from which the predictive distribution is

1. Data model: $[Z|Y,\theta]$
2. Process model: $[Y|\theta]$,

$$[Y|Z,\theta] = \frac{[Z|Y,\theta][Y|\theta]}{[Z|\theta]}, \tag{3.4}$$

where $[Z|\theta] = \int [Z|Y,\theta][Y|\theta]dY$. One can replace (3.4) with $[Y|Z,\hat{\theta}]$, where $\hat{\theta}$ is an estimator of θ, which can also be estimated from an independent study.

Depending on the complexity of the data model and the process model, sometimes it is possible to consider the frequentist statistical estimation for the parameters θ. In the context of EHM, common approaches include maximum likelihood estimation, the Expectation-Maximisation (EM) algorithm, conditional likelihood and pseudo-likelihood methods (Demidenko, 2013).

3.2 Frequentist approach

The HM introduces data Z, process Y and parameters θ; frequentist approach has only data Z and parameters θ, i.e., inference is based on the *likelihood*, $[Z|\theta]$. An estimation of the parameter θ, given the observed data Z could be found by maximising the likelihood or equivalently the log-likelihood function, that is finding $\hat{\theta}$ such as

$$\log([\hat{\theta}|Z]) > \log([\theta|Z]) \quad \forall\theta, \tag{3.5}$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE). On the other hand, Bayesians base their inferences on the *posterior* distribution, $[\theta|Z]$, which requires both a likelihood and *prior*, $[\theta]$, to be specified.

In the HM approach, there are some cases where models can be simplified due to practical concerns. The computational issues given formulation are limiting, which can lead to a modification of the model. Such concerns are not limited to HM inference, it can happen in any complicated modelling scenarios, see Casella & Berger (2002).

3.3 Bayesian hierarchical models (BHM)

A BHM is often useful for complex-modelling situations, because the parameters themselves can exhibit complex spatio-temporal structure. Or they could depend on other covariates and hence could be considered as processes by itself. The BHM approach al-

lows complex processes to be modelled by the hierarchy, but at each level the conditional-probability model can be simple. Other frameworks such as Machine learning uses a similar approach with its deep models. A potential advantage of the BHM approach is that it provides a unified probabilistic framework that allows one to account for uncertainty in data, model, and parameters.

Consider Z, Y and θ generic random variables; Z can be thought as data, Y as a hidden process and θ as unknown parameters. In general, these random quantities can be complicated, a spatial statistical mapping of a region's air quality in a given day, for instance. The three levels of a Bayesian hierarchical models are

1. Data model: $[Z|Y,\theta]$
2. Process model: $[Y|\theta]$
3. Parameter model: $[\theta]$

Note that the join distribution can be decomposed recursively.

$$\begin{aligned}[Z,Y,\theta] &= [Z,Y|\theta][\theta] \\ &= [Z|Y,\theta][Y|\theta][\theta]. \end{aligned} \tag{3.6}$$

The conditional distribution of Y and θ, given the data Z, which is called the *posterior distribution*, can be obtained from the application of Bayes' Rule (3.3) to the hierarchical model,

$$\begin{aligned}[Y,\theta|Z] &= \frac{[Z|Y,\theta][Y,\theta]}{\int\int[Z|Y,\theta][Y,\theta]dY\,d\theta} \\ &= \frac{[Z|Y,\theta][Y|\theta][\theta]}{\int\int[Z|Y,\theta][Y,\theta]dY\,d\theta} \\ &= \frac{[Z|Y,\theta][Y|\theta][\theta]}{[Z]}. \end{aligned} \tag{3.7}$$

All inference on Y and θ in the BHM depends on this distribution within the Bayesian framework. Numerical evaluation of this predictive (or posterior) distribution is needed, since the BHM's normalising constant cannot generally be calculated in closed form. An example of the BHM approach can be found in Wikle *et al.* (1998), where a Markov chain Monte Carlo framework is applied to an atmospheric data set of monthly maximum temperature.

There are some difficulties related to building Bayesian hierarchical models and then

carrying out valid inferences. A common criticism of Bayesian methods is that they require "subjective" specification of prior information on the parameters. However, one can notice that there is also subjectivity in the specification of the likelihood in frequentist models.

Computational issues in the Bayesian implementation of hierarchical spatial models are covered in the book Diggle & Ribeiro (2007). The dominant method is a MCMC approach. In cases where parameters are random, one should be careful about computational considerations such as convergence and efficiency. In addition, a prior distribution on parameters must be assumed. For instance, Gaussian-process usually include parameter associated with the variance, spatial dependence and smoothness. In many Bayesian analysis, the smoothness parameter is assumed to be known. In this case, one is focused on the variance and spatial-dependence parameters. For example, inverse gamma or discrete uniform priors are often chosen for the spatial-dependence parameter. On the other hand, the use of noninformative priors for variance components can be problematic (Gelman, 2006).

3.3.1 Bayesian Inference

The Bayesian approach considers that uncertainty can be described in terms of probability distributions. However, in many cases it implies a high number of parameters involved into a model. What follows is a brief introduction to some of the concepts involved in performing Bayesian analysis.

For a more information, see Gelman *et al.* (2013), Bernardo & Smith (1994). In assigning the prior distributions of the parameters, θ, the Bayesian perspective gives a natural framework for dealing with hierarchical models. This approach incorporates the uncertainty in the estimates of the parameters.

One could be interested in making conclusions about the unknown parameters, θ during statistical analyses. In other contexts the interest is in predicting values, Z, of the response variable for particular values of the explanatory variables. When data, Y, is observed, such inference will be expressed as $\pi(\theta|Y)$ and $\pi(Z|y)$ for the cases of parameter estimation and prediction respectively.

Let us considers the *prior distribution,* $\pi(\theta)$ and the density function of y given the value of the parameters, $\pi(y|\theta)$. Then, the joint probability distribution relating the parameters, and the observed values, $\pi(\theta, y)$, can be expressed as a product of two

densities.

$$\pi(\theta, y) = \pi(y|\theta)\pi(\theta). \tag{3.8}$$

As a result of using the Bayes' theorem, the *posterior density,* $\pi(\theta|y)$, representing the updated beliefs about θ after the data has been observed, can be obtained

$$\pi(\theta|y) = \frac{\pi(\theta, y)}{\pi(y)} = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \tag{3.9}$$

where $\pi(y) = \int_\theta \pi(y|\theta)\pi(\theta)d\theta$, over all the possible values of θ. This is true for the case of a continuous parameter, θ.

Assuming that y has been observed, then, the denominator, $\pi(y)$, does not depend on any unknown quantities. It is considered fixed and known as the *normalising constant.* The equation 3.9 can be simplified to

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta). \tag{3.10}$$

The prior distribution allows knowledge from previous studies or experiments to be incorporated. However, an increased level of complexity can be introduced into the calculations. In some circumstances there are actual data available that can be used to assign priors. But, commonly decisions have to be based on knowledge of other sources.

If one is interested in prediction, there are two cases to consider, the first one is before data y has been observed. The *marginal distribution* of y, $\pi(y)$, gives a distribution function for y. It is known as *prior predictive distribution* and does not depend on any previous observations.

In the second case, when the data, y, have been observed, a predicted value, Z, is obtained, as follows

$$\begin{aligned} \pi(Z|y) &= \int \pi(Z, \theta|y)d\theta \\ &= \int \pi(Z|y, \theta)\pi(\theta|y)d\theta \\ &= \int \pi(Z|\theta)\pi(\theta|y)d\theta \end{aligned} \tag{3.11}$$

It is assumed that the predicted value, Z, is conditionally independent of the observed data y given θ.

If a point estimate from the posterior distribution of the unknown parameters is required, then the posterior mean can be calculated by

$$\overline{\theta} = E(\theta|y) = \int \theta\pi(\theta|y)d\theta. \tag{3.12}$$

However, this is not necessarily an easy task. Alternatively, if samples can be taken from the posterior distribution, using, for example Gibbs sampling (see Section 3.4.1), then the mean and median together with credible intervals can be obtained. A $100(1-\alpha)\%$ *Credible Interval*, I, for this point estimate can be found by choosing a and b, with the set $I = (a, b)$ such that for a given α, $\pi(\theta \in I|y) = \int_b^a \pi(\theta|y)d\theta = 1 - \alpha$.

3.4 Computation

In Bayesian inference, in most cases, the posterior distribution of the parameter θ is analytically intractable. This means that it is not possible to get closed form summaries of the posterior distribution such as its mean and variance of a particular parameter. Simulation-based inference is a standard solution, the posterior distribution can be formulated in terms of simulation from an periodic and irreducible Markov chain.

3.4.1 Markov Chain Monte Carlo Methods

One approach to finding the posterior probabilities, which avoids the problems of integration, is to use simulation techniques, such as Markov Chain Monte Carlo (MCMC). For further information on MCMC see Robert & Casella (2004) and Gamerman & Lopes (2006). The Markov chain is constructed so that the stationary distribution is the posterior distribution. Two such Markov chains are the Metropolis-Hastings algorithm and the Gibbs sampler, which are now briefly described.

If one is able to draw $\theta_1, \ldots, \theta_N$, independent and identically distributed samples from the posterior distribution of θ, the standard Monte Carlo method can be considered. The mean of any function $g(\theta)$ having finite posterior expectation can be approximated by a sample average

$$E_\pi(g(\theta)) \approx N^{-1} \sum_{j=1}^{N} g(\theta_j). \tag{3.13}$$

We should notice that, in general, independent samples are not easy to obtain, and 3.13 holds for certain Markov chains. *Markov chain Monte Carlo* (MCMC) is the generic algorithm to perform the numerical analysis required by Bayesian data analysis.

Let us consider an irreducible, aperiodic and recurrent Markov chain $(\theta^t)_{t\geq 1}$, follow-

ing the invariant distribution π. It can be shown that for every initial value θ_1, the distribution of θ_t tends to π as t increases to infinity. Then, for M sufficiently large, $\theta_{M+1}, \ldots, \theta_{M+N}$ are all approximately distributed according to π. Jointly, these draws have statistical properties similar to an independent sample from π. By the law of large numbers, $E_\pi(\theta)$ can be approximated by

$$E_\pi(g(\theta)) \approx N^{-1} \sum_{j=1}^{N} g(\theta_{M+j}). \tag{3.14}$$

In practice, it is important to determine M, these first realisations correspond to the "burn-in" period and they should be discarded to calculate 3.14. In the remainder of this section, the most popular MCMC algorithms for simulating from a given distribution π are briefly explained.

3.4.2 Gibbs sampler

Let us suppose that the unknown parameter is multidimensional, then the distribution is multivariate. In this case, $\theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ and the target density is $\pi(\theta) = \pi(\theta^{(1)}, \ldots, \theta^{(k)})$. The Gibbs sampler starts from an arbitrary point $\theta_0 = (\theta_0^{(1)}, \ldots, \theta_0^{(k)})$ in the parameter space and updates one component at a time by drawing $\theta^{(i)}$, for $i = 1, \ldots, k$.

Algorithm: Gibbs sampler
Initialise the starting point: $\theta_0 = (\theta_0^{(1)}, \ldots, \theta_0^{(k)})$.
for $j = 1$ to N **do**
1. Obtain $\theta_j^{(1)}$ from $\pi(\theta^{(1)} | \theta^{(2)} = \theta_{j-1}^{(2)}, \ldots, \theta^{(k)} = \theta_{j-1}^{(k)})$
2. Obtain $\theta_j^{(2)}$ from $\pi(\theta^{(2)} | \theta^{(1)} = \theta_j^{(1)}, \theta^{(3)} = \theta_{j-1}^{(3)}, \ldots, \theta^{(k)} = \theta_{j-1}^{(k)})$

$\vdots$

k. Obtain $\theta_j^{(k)}$ from $\pi(\theta^{(2)} | \theta^{(1)} = \theta_j^{(1)}, \ldots, \theta^{(k-1)} = \theta_{j-1}^{(k-1)})$

end for

In some applications, the algorithm 3.4.2 is applicable when one or more of the components $\theta^{(i)}$ is itself multidimensional. The Gibbs sampler updates in turn *blocks* of components of θ, drawing from their conditional distribution, given all the remaining components.

3.4.3 Metropolis–Hastings algorithm

The Metropolis-Hastings is a general algorithm that allows us to generate the next state of the chain from an arbitrary distribution. It starts with the target density π. Suppose that the chain is currently at θ, then the *proposal* is obtained from a density $q(\theta|\cdot)$. In practice, this algorithm can be implemented when q is easy to simulate. The proposal $\tilde{\theta}$ is accepted as the new state of the chain with probability

$$\alpha(\theta, \tilde{\theta}) = \min\left\{1, \frac{\pi(\tilde{\theta})q(\tilde{\theta}, \theta)}{\pi(\theta)q(\theta, \tilde{\theta})}\right\}. \tag{3.15}$$

Algorithm: Metropolis–Hastings
Initialise the starting point: θ_0.
for $j = 1$ to N **do**
1. Obtain $\tilde{\theta_j}$ from $q(\theta_{j-1}, \cdot)$
2. Calculate $\alpha = \alpha(\theta_{j-1}, \tilde{\theta_j})$ de acording with (3.15)
3. Generate an independent random variable $U_j \sim \mathcal{B}ernoulli(\alpha)$

k. If $U_j = 1$ set $\theta_j = \tilde{\theta_j}$, otherwise set $\theta_j = \theta_{j-1}$
end for

If the proposal is rejected, the chain stays in the current state θ. The involved steps are shown in 3.4.3, assuming the chain starts at an arbitrary value θ_0. The selection of q is crucial, in particular if q is symmetric the expression 3.15 can be simplified as,

$$\alpha(\theta, \tilde{\theta}) = \min\left\{1, \frac{\pi(\tilde{\theta})}{\pi(\theta)}\right\}. \tag{3.16}$$

The Gibbs sampler and Metropolis–Hastings algorithm can work simultaneously in Markov chain simulation, they can be combined and used together. Within a Gibbs sampling simulation, it could be unfeasible to sample from one or more conditional distributions. Suppose for example that $\pi(\theta_1|\theta_2)$ does not have a standard form and is difficult to simulate from. In this situation one can, instead of generating θ_1 from $\pi(\theta_1|\theta_2)$, update θ_1 using a Metropolis–Hastings step. The invariant distribution of the Markov chain is not altered.

3.4.4 Analytical approximations

In some cases it is possible to find analytical expressions for posterior distributions. However, in most of the situations it is not possible to evaluate the posterior model

probabilities. In that case, it is necessary to calculate integrals of the form $p(y) = \int p(y|\theta)p(\theta)d\theta$. One alternative is considering the Bayesian statistics and computational techniques (MCMC) that were discussed in previous section, see Gelman *et al.* (2013) and Robert & Casella (2004).

Another approach is to use an analytic approximation such as Laplace's method, which is based on the Taylor series expansion of a real valued function $f(u)$

$$\int e^{f(u)}du \approx (2p)^{r/2}|H|^{1/2}\exp\{f(u^*)\} \tag{3.17}$$

where r is the length of the vector u, u^* is the value of u at which f attains its maximum and H is minus the inverse Hessian information of f evaluated at u^*.

The integrated nested Laplace approximation (INLA) for approximate Bayesian inference was developed by Rue *et al.* (2009). It is an alternative to traditional MCMC methods. A wide variate of models of INLA can be found in Krainski *et al.* (2018).

Now, a general description of INLA is given, see Krainski *et al.* (2018). Let us consider a a vector of observed variables whose distribution is in the exponential family, $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. The mean, μ_i is linked to the linear predictor η_i using an appropriate link function. This linear predictor can include terms on covariates (fixed effects) and different types of random effects. The vector of all latent effects will be denoted by $\mathbf{x}$. And, the distribution of $\mathbf{y}$ will depend on some vector of hyperparameters, $\boldsymbol{\theta}_1$.

The distribution of the vector of latent effects $\mathbf{x}$ is assumed to be Gaussian Markov random field (GMRF). The GMRF will have a zero mean and precision matrix $\mathbf{Q}_{\boldsymbol{\theta}_2}$, with $\boldsymbol{\theta}_2$ a vector of hyperparameters. The vector of all hyperparameters in the model will be denoted by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. If the observations are assumed to be independent given the vector of latent effects and the hyperparameters. Then, the the likelihood is given by

$$\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{i\in\mathcal{I}} \pi(y_i|\eta_i, \boldsymbol{\theta}).$$

Here, ν_i is the latent linear predictor and the set $\mathcal{I}$ contains indices for all observed values of $\mathbf{y}$. The idea of the INLA methodology is to approximate the posterior marginals of the model effects and hyperparameters. This is done by exploiting the computational properties of GMRF and the Laplace approximation for multidimensional integration.

The joint posterior distribution of the effects and hyperparameters is given by

$$\begin{aligned}\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i\in\mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\mathbf{Q}_{\boldsymbol{\theta}}|^{1/2}\exp\Big\{ -\frac{1}{2}\mathbf{x}'\mathbf{Q}_{\boldsymbol{\theta}}\mathbf{x} + \sum_{i\in\mathcal{I}} \log(\pi(y_i|x_i, \boldsymbol{\theta}))\Big\}.\end{aligned}$$

Here, $\mathbf{Q}_{\boldsymbol{\theta}}$ is the precision matrix of the latent effects and $|\mathbf{Q}_{\boldsymbol{\theta}}|$ denotes the determinant of that precision matrix. In particular, when $i \in \mathcal{I}$ then $x_i = \eta_i$. The marginal distributions for the latent effects and hyperparameters can obtained by

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

and

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}.$$

where integration is done over the space of the hyperparameters. A good approximation to the joint posterior distribution of the hyperparameters is required. In Rue *et al.* (2009), $\pi(\boldsymbol{\theta}|\mathbf{y})$ is approximated and denoted by $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. Then, the posterior marginal of the latent parameter x_i can be approximated by

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k, \mathbf{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) \times \Delta_k,$$

where Δ_k are the weights associated with a vector of values $\boldsymbol{\theta}_k$ of the hyperparameters in a grid.

3.5 Discussion

In this chapter, we introduce the concept of a hierarchical model, which is comprised of a data model which expresses the distribution of the data given a latent (unobserved) process. This latent process is represented by a statistical model that describes the underlying dependencies within the process that generates the data. In the case of a Bayesian model, a third level, the parameter level, assigns (prior) distributions to all of the parameters in the model. We summarise different approaches to performing inference, including empirical, frequentist and Bayesian methods. For the latter, we describe different methods for computation, including Markov Chain Monte Carlo simulation (Gibbs Sampling and Metropolis-Hastings) and the use of analytical approximations. The latter, through the use of integrated nested Laplace approximation (INLA) is used for the Bayesian Hierarchical Spatio-Temporal Models (BHSTM) presented in Chapter

7 and offers a computationally efficient way of implementing complex spatio-temporal Bayesian models.

Chapter 4

Spatio-temporal modelling

We now present a brief summary of spatio-temporal statistical modelling, and the background of the concepts and methods that are used in this thesis. One can distinguish between the underlying spatio-temporal process and the process in which measurements are made. In particular, let $\{Y(\mathbf{s};t) : \mathbf{s} \in D_{\mathbf{s}}, t \in D_t\}$ denote a spatio-temporal random process that is a statistical model for a phenomenon involving the spatio-temporal index set $D_{\mathbf{s}} \times D_t$. For instance, $Y(\mathbf{s};t)$ can be measured air pollution at geographical coordinates $\mathbf{s}$ = (latitude,longitude) at time t. Conceptually, $D_{\mathbf{s}}$ = plane and $D_t = \mathbb{R}$. This process is not directly measurable, but the realisations $Y(\mathbf{s};t)$ can be obtained by taking measurements, possibly with an error.

The mean and covariance function can be estimated from the data, they also can be used to describe the spatial trend and for prediction at unsampled locations. Spatio-temporal covariance functions offer an informative summary of a random process on $D_{\mathbf{s}} \times D_t$, indeed, the covariance function can be thought as a characterisation of the spatio-temporal process $Y(\cdot;\cdot)$. The covariance function and the semivariogram are both functions that summarise the relation in terms of distance and differences in time.

One should note that for most of the real-world problems in spatio-temporal statistics, there is a relatively small number of monitoring sites for a large area of interest. Due to the sparsity of such sites, one should be able to interpolate or extrapolate the measurements at these sites in order to obtain a spatio-temporal predictions. There are different interpolation methods, one of the most simple consist of a weighted average of all spatial and temporal points within a certain neighbourhood of a chosen point, $(\mathbf{s}_0, t_0)$. One of these deterministic methods is discussed in section 2.2. Another popular

method for spatio temporal prediction is *kriging*, this is discussed in section 6.1. A classic reference for the purely spatial kriging is Cressie (1993). For a further discussion about spatio-temporal kriging see Le & Zidek (2006); Cressie & Wikle (2011).

4.1 Covariance Functions

A covariance function is a *nonnegative-definite function* and vice versa. This is why spatio-temporal statistics is interested in the study of nonnegative-definite functions in space and time.

Definition 4.1.1. A function $\{f(\mathbf{u}, \mathbf{v}) : \mathbf{u}, \mathbf{v} \in D\}$ defined on $D \times D$ is said to be *nonnegative-definite*, if for any complex numbers $\{a_i : i = 1, \ldots, m\}$ any $\{\mathbf{u}_i : i = 1, \ldots, m\}$ in D, and any m, one has

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i \bar{a}_j f(\mathbf{u}_i, \mathbf{u}_j) \geq 0, \tag{4.1}$$

where it is recalled that $\bar{a}$ denotes the complex conjugate of a.

If the inequality above is strictly positive one says that $f(\cdot, \cdot)$ is *positive-definite* whenever $a \equiv (a_1, \ldots, a_m)'$ is a nonzero vector.

If D is a subset of integers $\mathbb{Z}$, f is a temporal covariance function for a discrete-index time series. On the other hand, if $D \in \mathbb{R}^d$, then f is a spatial covariance function for a continuous-index spatial process. In the spatial-temporal setting, D could be a subset of $\mathbb{R}^d \times \{\ldots, -1, 0, 1, \ldots\}$ or $\mathbb{R}^d \times \mathbb{R}$. Then, a spatial covariance function in $\mathbb{R}^{d+1}$ could be used as a spatio-temporal covariance function in $\mathbb{R}^d \times \mathbb{R}$. One can write $\mathbf{u} = (\mathbf{s}; t)$ and $f(\mathbf{u}_i, \mathbf{u}_j) \equiv f((\mathbf{s}_i; t_i), (\mathbf{s}_j, t_j))$, to obtain a spatio-temporal covariance function.

The equation 4.1 can be rewritten in terms of $C(\cdot; \cdot)$ which is defined on $\mathbb{R}^d \times \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i \bar{a}_j C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \geq 0, \tag{4.2}$$

for any $\{a_i\}$, any $(\mathbf{s}_i, t_i)$ and any m.

4.1.1 Stationarity in Space and Time

Definition 4.1.2. It is said that f is a *stationary* spatio-temporal covariance function on $\mathbb{R}^d \times \mathbb{R}$, if it is a *nonnegative-definite function* and can be written as

$$f((\mathbf{s}_i; t_j), (\mathbf{s}_k, t_l)) = C(\mathbf{s}_i - \mathbf{s}_k; t_j - t_l), \quad \mathbf{s}_i, \mathbf{s}_k \in \mathbb{R}^d, t_j, t_l \in \mathbb{R}. \tag{4.3}$$

If a random process $Y(\cdot;\cdot)$ has a constant expectation and a stationary covariance function, $C_Y(\mathbf{h};\tau)$, then it is said to be *second-order (or weakly) stationary*. Strong stationary of $Y(\cdot;\cdot)$ corresponds to the equivalence of the two probability measures defining the random process $Y(\cdot;\cdot)$ and $Y(\cdot + \mathbf{h}; \cdot + \tau)$, respectively, for all $\mathbf{h} \in \mathbb{R}^d$ and all $\tau \in \mathbb{R}$. When there is no ambiguity, one can omit dependence on Y and write, $C(\mathbf{h};\tau)$.

The *stationary spatio-temporal correlation function* associated with $C(\cdot;\cdot)$ is

$$\rho(\mathbf{h};\tau) \equiv C(\mathbf{h};\tau)/C(\mathbf{0};\tau) \qquad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}. \tag{4.4}$$

The correlation function 4.4 represents spatio-temporal dependence in a continuous-space ($\mathbb{R}^d$), continuous time ($\mathbb{R}$) model.

Stationarity of the covariance function can be considered separately for space and time. *Spatial stationarity* of the covariance function corresponds to

$$\text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) \equiv C(\mathbf{s}_i - \mathbf{s}_k; t_j, t_l) \tag{4.5}$$

and *temporal stationarity* of the covariance function corresponds to

$$\text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) \equiv C(\mathbf{s}_i, \mathbf{s}_k; t_j - t_l) \tag{4.6}$$

The *spatio-temporal stationarity* of the covariance function corresponds to

$$\text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) \equiv C(\mathbf{s}_i - \mathbf{s}_k; t_j - t_l). \tag{4.7}$$

Similar to spatial stationarity, *spatial isotropy* corresponds to

$$\text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) \equiv C(||\mathbf{s}_i - \mathbf{s}_k||; t_j, t_l). \tag{4.8}$$

Spatio-temporal kriging predictions require that one knows the spatio-temporal covariances between the hidden random process evaluated at any two locations in space and

time. It is important to note that not any function can be used as a covariance function, a formal definition was given in 4.1.1.

In practice, for classical-kriging implementations, second-order stationarity is assumed; recall definition in 4.1.2. There are some advantages of the second-order stationarity assumption, such as it allows us to give more parsimonious parameterisations of the covariance function. It also helps to estimate parameters, say $\boldsymbol{\theta}$, involved in the covariance function. The parameterisation of these covariance functions is one of the most challenging problems in spatio-temporal statistics.

4.2 Separability and Full Symmetry

Definition 4.2.1. A random process $Y(\cdot;\cdot)$ is said to have a *separable spatio-temporal covariance function* if, for all $\mathbf{s}_i, \mathbf{s}_k \in \mathbb{R}^d, t_j, t_l \in \mathbb{R}$, it is obtained

$$\operatorname{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) = C^{(\mathbf{s})}(\mathbf{s}_i, \mathbf{s}_k) \cdot C^{(t)}(t_j, t_l) \tag{4.9}$$

where $C^{(\mathbf{s})}$ and $C^{(t)}$ are spatial and temporal covariances functions, respectively.

A simple case of spatio-temporal covariance functions can be obtained as a consequence of equation 4.9. When $C^{(\mathbf{s})}$ and $C^{(t)}$ are spatially and temporally stationary, respectively, then

$$C(\mathbf{h};\tau) = C^{(\mathbf{s})}(\mathbf{h}) \cdot C^{(t)}(\tau), \qquad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}. \tag{4.10}$$

Separability implies that the spatio-temporal correlation function $\rho(\cdot,\cdot)$ given by equation 4.4 satisfies

$$\rho(\mathbf{h};\tau) = \rho^{(\mathbf{s})}(\mathbf{h};0) \cdot \rho^{(t)}(0;\tau) \qquad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}, \tag{4.11}$$

where $\rho^{(\mathbf{s})}(\mathbf{h};0)$ and $\rho^{(t)}(0;\tau)$ are the corresponding marginal spatial and temporal correlation functions, respectively. The converse is true, and then equation 4.11 is a characterisation of separability in second-order stationary process. This result allows one to make a visual inspection of separability. One can compare the contours of $C(\mathbf{h};\tau)$ with those of $C(\mathbf{h};0) \cdot C(0;\tau)$

Definition 4.2.2. A random process $Y(\cdot;\cdot)$ is said to have a *fully symmetric spatio-*

temporal covariance function if, for all $\mathbf{s}_i, \mathbf{s}_k \in \mathbb{R}^d, t_j, t_l \in \mathbb{R}$,

$$\text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)) = \text{cov}(Y(\mathbf{s}_i; t_l), Y(\mathbf{s}_k; t_j). \tag{4.12}$$

Considering such covariances to model spatio-temporal dependence is not always reasonable for real-world processes. One should notice that in general, separable covariance functions are always fully symmetric, while the converse is not true.

4.2.1 Examples of Separable (in Space and Time) Covariance Functions

Separable classes of spatio-temporal covariance functions have been used in spatio-temporal modelling because they offer a convenient way to verify validity. These classes have been defined in 4.2.1, recall the equation 4.10,

$$C(\mathbf{h}; \tau) = C^{(\mathbf{s})}(\mathbf{h}) \cdot C^{(t)}(\tau), \qquad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R} \tag{4.13}$$

which is valid if both the spatial covariance function and the temporal covariance function are valid.

There are many classes of valid spatial and valid temporal covariance functions in the literature. For example the Matérn, the power exponential, and Gaussian classes.

The Matérn family of correlation functions is commonly used; it is written as

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right), \qquad d > 0, \tag{4.14}$$

where $\Gamma(\cdot)$ is the gamma function, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, and ρ, $\nu > 0$ are parameters of the covariance Abramowitz & Stegun (1972).

Some examples of these classes of spatial covariance functions can be defined as the distance between two locations, i.e., $d \equiv ||\mathbf{h}||$ where $\mathbf{h} \in \mathbb{R}^d$ and $\rho \equiv a_\mathbf{s}$, see table 4.1.

In all cases, $\sigma^2_\mathbf{s}$ is the variance parameter and $a_\mathbf{s}$ is the spatial-dependence (or scale) parameter in units of distance. The larger $a_\mathbf{s}$ is, the more dependent the spatial process is. In figure 4-1 one can observe the different behaviour of the exponential, spherical and Gaussian models with the same variance ($\sigma^2_\mathbf{s} = 1$) and also the same scale parameter ($a_\mathbf{s} = 1/3$).

In a similar way, a valid temporal covariance function can be defined, such as the

Model	Spatial covariance function	Parameters
Exponential	$C^{(\mathbf{s})}(\mathbf{h}) = \sigma^2_{\mathbf{s}} \exp\left(-\frac{\|\|\mathbf{h}\|\|}{a_{\mathbf{s}}}\right)$	$0 < \|\|\mathbf{h}\|\|$
Spherical	$C^{(\mathbf{s})}(\mathbf{h}) = \sigma^2_{\mathbf{s}} \left(1 - \frac{3}{2}\frac{\|\|\mathbf{h}\|\|}{a_{\mathbf{s}}} + \frac{1}{2}\left(\frac{\|\|\mathbf{h}\|\|}{a_{\mathbf{s}}}\right)^3\right)$	$0 < \|\|\mathbf{h}\|\| < \frac{1}{a_{\mathbf{s}}}$
Gaussian	$C^{(\mathbf{s})}(\mathbf{h}) = \sigma^2_{\mathbf{s}} \exp\left(-\frac{\|\|\mathbf{h}\|\|^2}{a^2_{\mathbf{s}}}\right)$	$0 < \|\|\mathbf{h}\|\|$
Powered exponential	$C^{(\mathbf{s})}(\mathbf{h}) = \sigma^2_{\mathbf{s}} \exp\left(-\left(\frac{\|\|\mathbf{h}\|\|}{a_{\mathbf{s}}}\right)^p\right)$	$0 < \|\|\mathbf{h}\|\|,\ 0 < p$

Table 4.1: Spatial covariance functions: exponential, spherical, Gaussian and power exponential.

temporal exponential covariance function,

$$C^{(t)}(\tau) = \sigma^2_t \exp\left(-\frac{|\tau|}{a_t}\right). \tag{4.15}$$

Another reason to consider separable models is that they can make computation easier. Let us consider the observations $Z(\mathbf{s}_{ij}, t_j)$, for the same $i = 1, \ldots, m_j = m$ locations at at each time point $j = 1, \ldots, T$. One can consider,

$$\mathbf{C}_Z = \mathbf{C}_Z^{(t)} \otimes \mathbf{C}_Z^{(\mathbf{s})}, \tag{4.16}$$

where $\otimes$ is the Kronecker product (see appendix). In this case, $\mathbf{C}_Z^{(t)}$ is the $T \times T$ is temporal covariance matrix, and $\mathbf{C}_Z^{(\mathbf{s})}$ is the $m \times m$ spatial covariance matrix. Using the property,

$$\mathbf{C}_Z^{-1} = (\mathbf{C}_Z^{(t)})^{-1} \otimes (\mathbf{C}_Z^{(\mathbf{s})})^{-1}, \tag{4.17}$$

instead of taking the inverse of the $mT \times mT$ matrix $\mathbf{C}_Z$, one can take the inverses of $T \times T$ and $m \times m$ matrices.

Due to the separability property, the temporal evolution of the process at a given spatial location does not depend directly on the temporal evolution at other locations. As a consequence, other classes of spatio-temporal covariance functions shall be developed; below are some of these approaches:

- sums-and-products formulation.
- construction by a spectral representation,

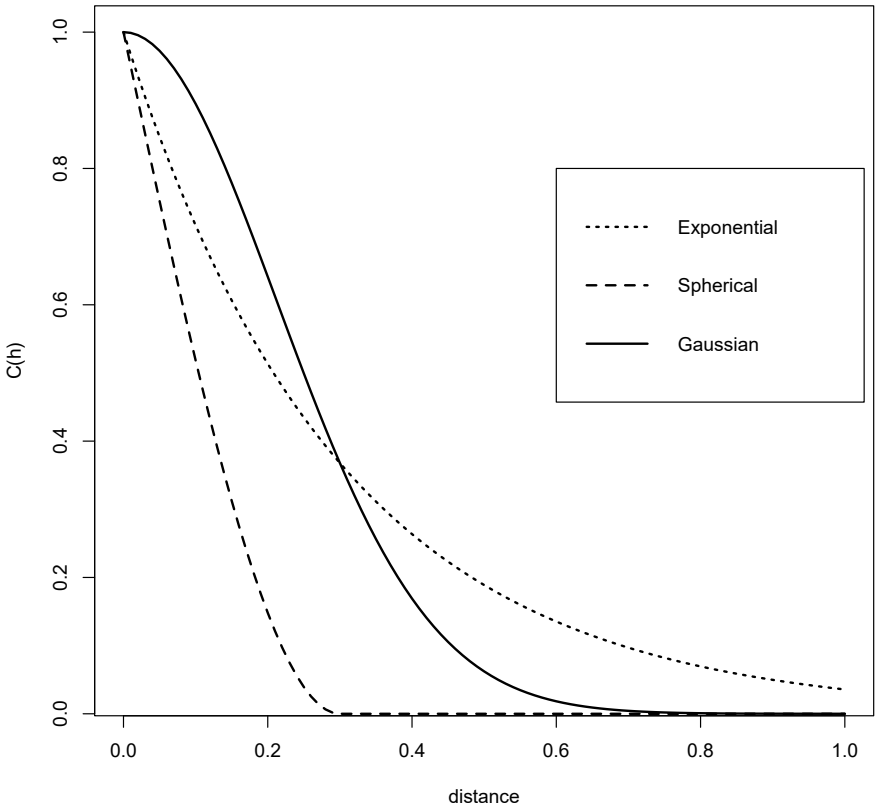

Figure 4-1: Exponential, spherical, and Gaussian models with $\sigma_{\mathbf{s}}^2 = 1$ and $a_{\mathbf{s}} = 1/3$.

- covariance functions from the solution of stochastic partial differential equations (SPDEs).

A general discussion about these approaches can be found in Cressie & Wikle (2011).

4.2.2 Sums and Products of Covariance Functions

Due to the fact that the sum and product of two nonnegative-definite functions are nonnegative, valid spatio-temporal covariance functions can be proposed:

$$C(\mathbf{h};\tau) \equiv p\, C^{(\mathbf{s}_1)}(\mathbf{h}) \cdot C^{(t_1)}(\tau) + q\, C^{(\mathbf{s}_2)}(\mathbf{h}) \cdot C^{(t_2)}(\tau), \tag{4.18}$$

where $p > 0, q > 0$, and $r > 0$, $C^{(\mathbf{s}_1)}$ and $C^{(\mathbf{s}_2)}$ are spatial covariances functions and $C^{(t_1)}$ and $C^{(t_2)}$ are temporal covariance functions.

There are other special cases; consider, for example the fully symmetric spatio-temporal covariance functions previously defined in equation 4.12. Although it could be appropriate under certain conditions, in general, the fully symmetric covariance function is not an appropriate choice.

4.3 Spatio-Temporal Semivariogram

There is another way to express dependence, related to the covariance function.

Definition 4.3.1. The spatio-temporal semivariogram of the process $Y(\cdot;\cdot)$ is defined to be

$$\gamma_Y(\mathbf{s}_i, \mathbf{s}_k; t_j, t_l) \equiv \frac{1}{2}\text{var}(Y(\mathbf{s}_i; t_j) - Y(\mathbf{s}_k; t_l)), \quad \mathbf{s}_i, \mathbf{s}_k \in \mathbb{R}^d, t_j, t_l \in \mathbb{R}. \tag{4.19}$$

If the covariance depends only on displacements in space and differences in time, then

$$\begin{aligned} \gamma_Y(\mathbf{h}; \tau) &= \frac{1}{2}\text{var}(Y(\mathbf{s}+\mathbf{h}; t+\tau) - Y(\mathbf{s}; t)) \\ &= C_Y(\mathbf{0}; 0) - \text{cov}(Y(\mathbf{s}+\mathbf{h}; t+\tau) - Y(\mathbf{s}; t)) \\ &= C_Y(\mathbf{0}; 0) - C_Y(\mathbf{h}; \tau), \end{aligned} \tag{4.20}$$

where $\mathbf{h} = \mathbf{s}_k - \mathbf{s}_i$ is a spatial lag and $\tau = t_l - t_j$ is a temporal lag.

Observe that equation 4.20 does not always hold. Even when γ_Y is defined as a function of spatial lag $\mathbf{h}$ and temporal τ, the stationary covariance function $C_Y(\mathbf{h}; \tau)$ is not necessarily defined.

Generally, the semivariogram is a non-decreasing monotone function, then the variability of the first increments of the random process increases with distance. The slope of the semivariogram represents the change in the dissimilarity of the values of the regionalised variable with distance. The limiting value of the semivariogram is called the sill (m), meanwhile the distance at which the sill is reached is called the range. It defines the threshold of spatial dependence, that is, the zone of influence of the random process.

In other words, the range is the distance beyond which the values of the regionalised variable have no spatial dependence. In figure 4-2 it can be observed that the larger the range, the larger the zone of influence of the random process.

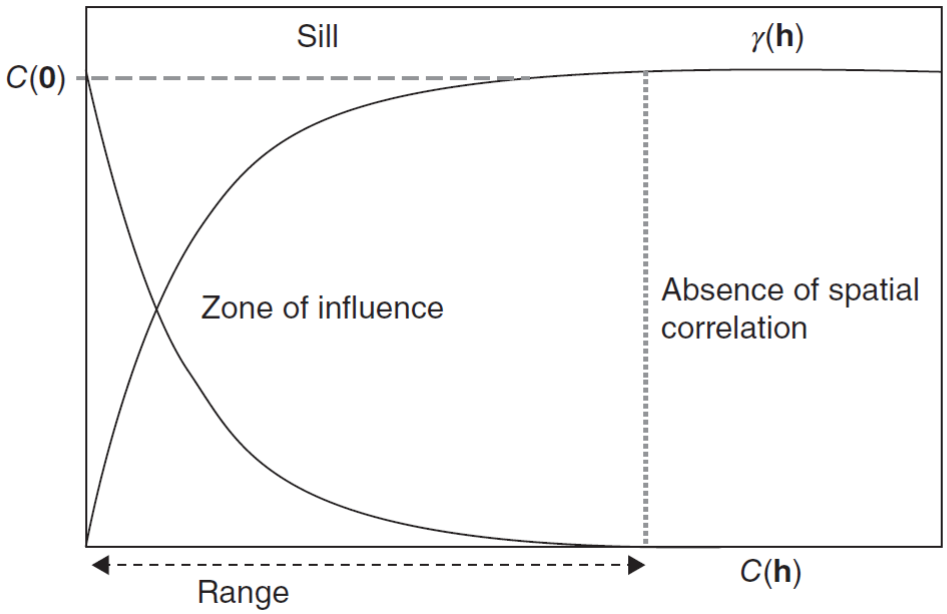

Figure 4-2: Example of semivariogram and its respective covariogram.

4.3.1 The Nugget Effect

The behaviour of the semivariogram at the origin is related to the continuity and degree of regularity of the random process. The more continuous and regular across space the rf, the smoother the more regular the behaviour of the semivariogram for short distances.

If the covariance function of the process is well defined, then the semivariogram is generally characterised by the nugget effect, the sill, and the partial sill, see figure 4-3.

The nugget effect is given by $\gamma_Y(\mathbf{h};\tau)$ when $\mathbf{h} \to 0$ and $\tau \to 0$. Although the semivariogram must be identically null at the origin, sometimes in practice this does not occur. It usually indicates that the regionalised variable is irregular, maybe discontinuous.

The sill is $\gamma_Y(\mathbf{h};\tau)$ when $\mathbf{h} \to \infty$ and $\tau \to \infty$ while the partial sill is the difference between the sill and the nugget effect. The limiting case of the nugget effect is the pure nugget effect. In particular, the semivariogram is constant for any given distance, indicating the absence of spatial dependence. A complete revision and examples of these concepts can be seen in Montero *et al.* (2015).

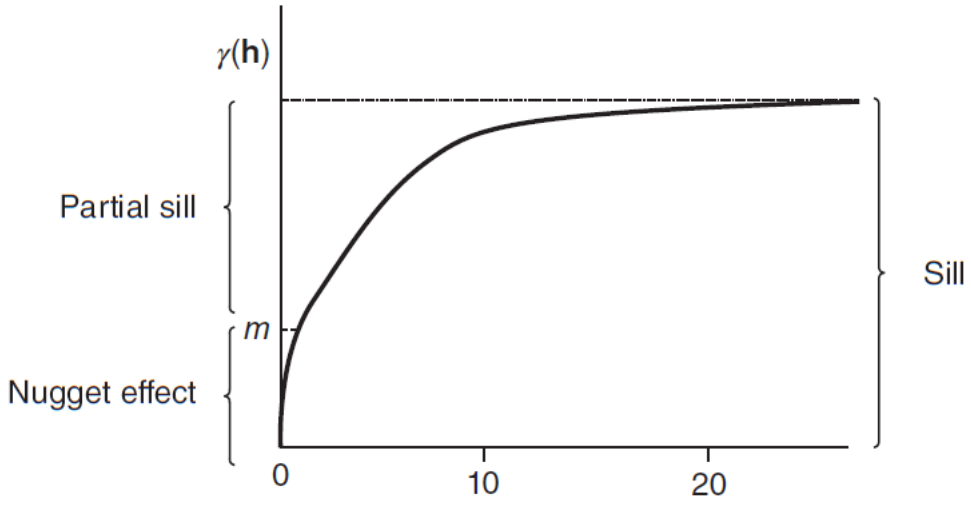

Figure 4-3: Example of the nugget effect, the sill, and the partial sill.

4.4 Application of empirical approaches to LAQN data

Visualisation of data is an important component of exploratory data analysis. In addition, in some circumstances it is necessary exploring spatio-temporal data in terms of summaries of first-order and second-order characteristics. Some visualisations of empirical means and empirical covariances, spatio-temporal covariograms and semivariograms are considered in this section. These concepts will be used to calculate the EOFs for the models described in section 5.1. The spatial mean that is specified in this section will be used to detrend the data set that will be considered later. Moreover, the empirical covariance matrices described here will be used to perform the DSTMs.

4.4.1 Empirical Spatial Means and Covariances

Assume that one has a realisation of a general spatio-temporal process, $Z(\mathbf{s}_i; t_j)$, for spatial locations $\{\mathbf{s}_i = 1, \ldots, m\}$ and times $\{t_j = 1, \ldots, T\}$.

Definition 4.4.1. The empirical spatial mean for location $\mathbf{s}_i$, $\hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_i)$, is defined by averaging over time

$$\hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_i) \equiv \frac{1}{T} \sum_{j=1}^{T} Z(\mathbf{s}_i; t_j). \tag{4.21}$$

Considering the mean of all spatial locations and if it is assumed that one has T observations at each locations, one has the spatial mean as an m-dimensional vector,

$\hat{\mu}_{z,\mathbf{s}}$, where

$$\hat{\boldsymbol{\mu}}_{z,\mathbf{s}} \equiv \begin{bmatrix} \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_1) \\ \vdots \\ \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_m) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_{j=1}^{T} Z(\mathbf{s}_1; t_j) \\ \vdots \\ \frac{1}{T}\sum_{j=1}^{T} Z(\mathbf{s}_m; t_j) \end{bmatrix} = \frac{1}{T}\sum_{j=1}^{T} \mathbf{Z}_{t,j}, \tag{4.22}$$

and $\mathbf{Z}_{t,j} \equiv (Z(\mathbf{s}_1, t_j)), \ldots, Z(\mathbf{s}_1, t_j))'$. This mean vector is a spatial quantity whose elements are indexed by their location.

The empirical spatial mean for NO_2, PM_{10} and $PM_{2.5}$ can be plotted as in figure 4-4 per each site. Based on these plots, one can say that highest levels of $PM_{2.5}$ can be observed in areas nearby Inner London. A high relation can be observed between the latitude and longitude of some locations and their respective $PM_{2.5}$ levels. In general, one can conclude that there is a trend in the empirical spatial mean of $PM_{2.5}$ with the longitude and latitude. The trend is similar to that one for PM_{10}, with the highest levels observed in the central area. In the case of NO_2, the behaviour of this pollutant is slightly different, observing a wider gap within the spatial means across London. Higher concentrations of NO_2 can also be found in Inner London.

One can average across space and plot the associated time series. In a similar way, one can define the empirical temporal mean across the space for t_j,$\hat{\mu}_{z,t}(t_j)$,

$$\hat{\mu}_{z,t}(t_j) \equiv \frac{1}{m}\sum_{i=1}^{m} Z(\mathbf{s}_i; t_j). \tag{4.23}$$

In figure 4-5, one can see the time series of NO_2, PM_{10} and $PM_{2.5}$ for each pollutant from the data set. The blue lines correspond to a station, and the empirical temporal mean, $\hat{\mu}_{z,t}(t_j))$, is the black line computed from equation 4.23. Time is in units of days, ranging from 01 August 2015 to 15 March 2016. A seasonal trend cannot be observed for all the pollutants, it could be due to the time series are not long enough. For NO_2 concentrations, a higher variability can be observed in contrast with PM levels for the same period of time. Visually, from figure 4-5 is not possible to conclude that PM_{10} and $PM_{2.5}$ levels have a seasonal behaviour. PM_{10} and $PM_{2.5}$ levels show a similar trend in time, however $PM_{2.5}$ has a larger variability. The last two plots of this figure confirm the high relation between these two pollutants.

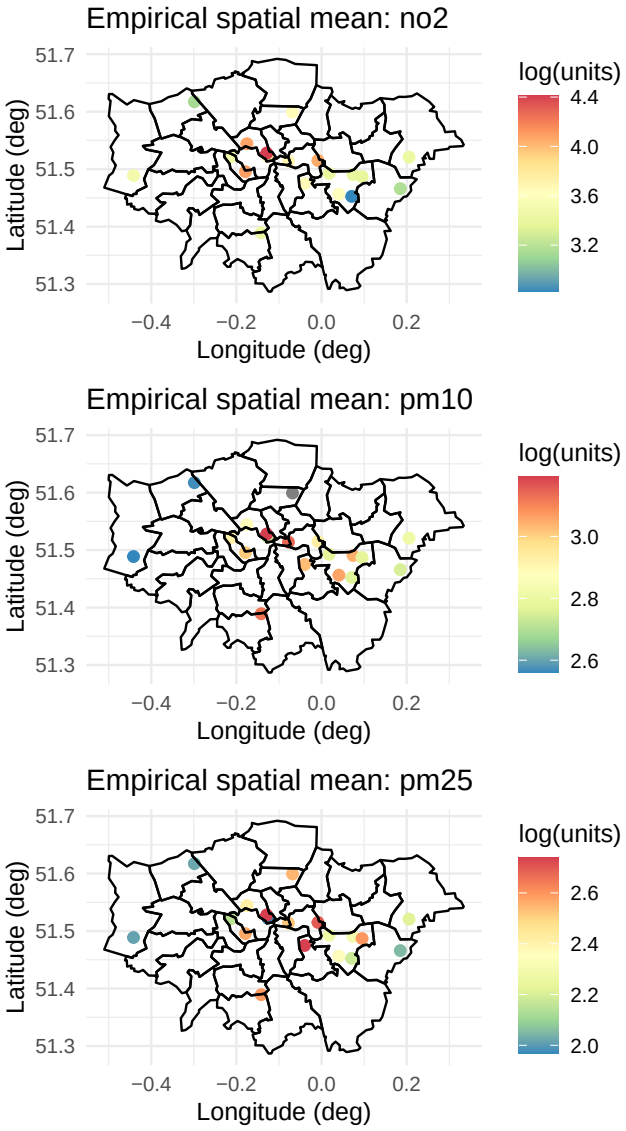

Figure 4-4: Empirical spatial mean in log-scale, $\hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_i)$, of NO_2, PM_{10} and $PM_{2.5}$. Summaries for the air quality network in Greater London area from 01 August 2015 to 15 March 2016.

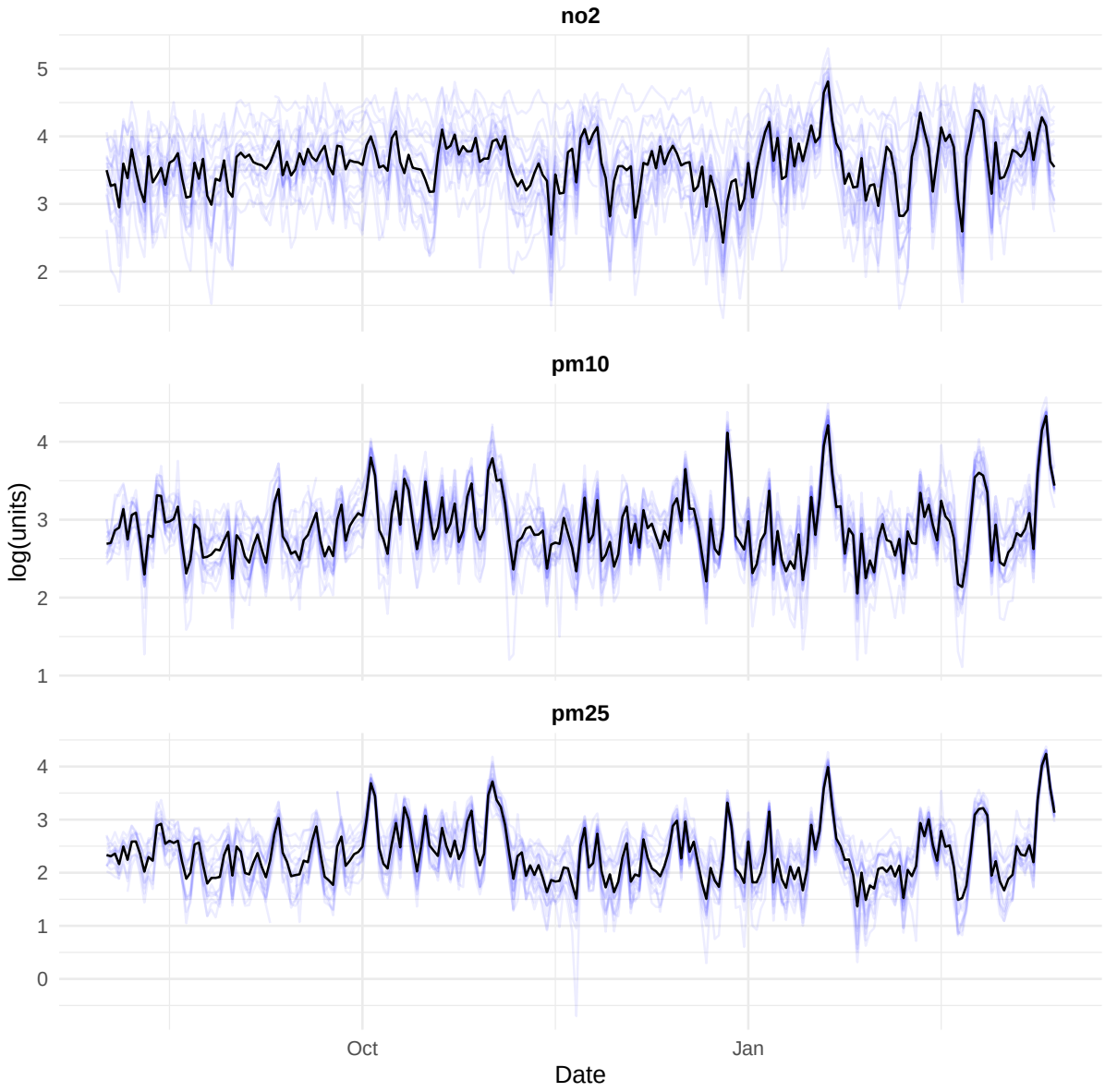

Figure 4-5: Empirical temporal mean in log-scale across the space for NO_2, PM_{10} and $PM_{2.5}$ (black line). Time series per pollutant for all sites (light blue lines). Summaries for the air quality network in Greater London area from 01 August 2015 to 15 March 2016.

It can also be informative to consider the empirical spatio-temporal covariance as a function of space and/or time.

Definition 4.4.2. The empirical lag-τ covariance between spatial locations $\mathbf{s}_i$ and $\mathbf{s}_k$ is given by

$$\hat{C}_z^{(\tau)}(\mathbf{s}_i, \mathbf{s}_k) \equiv \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} (Z(\mathbf{s}_i; t_j) - \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_i))(Z(\mathbf{s}_k; t_j - \tau) - \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_k)), \tag{4.24}$$

for $\tau = 0, 1, \ldots, T-1$.

It is also useful to consider an $m \times m$ empirical (averaged over time) lag-τ spatial covariance matrix given by

$$\hat{C}_z^{(\tau)} \equiv \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} (\mathbf{Z}_{tj} - \hat{\boldsymbol{\mu}}_{z,\mathbf{s}})(\mathbf{Z}_{tj-\tau} - \hat{\boldsymbol{\mu}}_{z,\mathbf{s}})'; \tag{4.25}$$

for $\tau = 0, 1, \ldots, T-1$, in which the (i,k)th element is given by equation 4.1. Due to locations in a two-dimensional space not having a natural ordering, finding a intuition from these matrices can be difficult.

4.4.2 Empirical Spatio-Temporal Covariograms and Semivariograms

It is needed to characterise the joint spatio-temporal dependence structure of a spatio-temporal process in order to perform optimal prediction. Thus, it is important to examine the empirical covariance function at various space and time lags, making the assumption that the first moment (mean), depends on space but not on time, and the second moment (covariance) depends only on spatial and temporal lag differences.

Definition 4.4.3. Then the empirical spatio-temporal covariogram for spatial lag $\mathbf{h}$ and time lag τ is given by

$$\hat{C}_z = \frac{1}{|N_s(\mathbf{h})|} \frac{1}{|N_t(\tau)|} \sum_{\mathbf{s}_i, \mathbf{s}_k \in N_{\mathbf{s}(\mathbf{h})}} \sum_{t_j, t_l \in N_t(\tau)} (Z(\mathbf{s}_i; t_j) - \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_i))(Z(\mathbf{s}_k; t_l) - \hat{\mu}_{z,\mathbf{s}}(\mathbf{s}_k)), \tag{4.26}$$

where it is recalled that $\hat{\mu}_{z,\mathbf{s}} = (1/T)\sum_{j=1}^{T} Z(\mathbf{s}_i, t_j)$, $N_{\mathbf{s}}(\mathbf{h})$ refers to the pairs of spatial locations with spatial lag within some tolerance of $\mathbf{h}$, $N_t(\tau)$ refers to the pairs of time points with time lag within some tolerance of τ, and $|N(\cdot)|$ refers to the cardinality

of the set $N(\cdot)$. Assuming isotropy, consider the lag only as a function of distance, $\mathbf{h} = ||h||$, where $||\cdot||$ is the Euclidean norm.

4.4.3 Empirical Semivariogram

One could be interested in the empirical spatio-temporal semivariogram, when the covariance only depends on the displacements in space and the time lags, it can be obtained from the equation 4.26 as

$$\hat{\gamma}_z(\mathbf{h};\tau) = \hat{C}_z(\mathbf{0};0) - \hat{C}_z(\mathbf{h};\tau). \tag{4.27}$$

If one has a constant spatial mean $\mu_{z,s}$, the equation 4.20 can be written as

$$\gamma_z(\mathbf{h};\tau) = \frac{1}{2}E(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s};t))^2. \tag{4.28}$$

Another estimate is

$$\hat{\gamma_z}(\mathbf{h};\tau) = \frac{1}{|N_s(\mathbf{h})|}\frac{1}{|N_t(\tau)|}\sum_{\mathbf{s}_i,\mathbf{s}_k\in N_{\mathbf{s}}(\mathbf{h})}\sum_{t_j,t_l\in N_t(\tau)}(Z(\mathbf{s}_i;t_j) - Z(\mathbf{s}_k;t_l))^2, \tag{4.29}$$

where the notation has used above in 4.26.

Thinking about spatio-temporal correlation simultaneously is not as simple as spatial or temporal correlation separately. To obtain an empirical spatio-temporal semivariogram that can be visually interpreted, one often groups together space-time lags into a set of space-time bins, this can be done trough the package `gstat`. The figure 4-6 shows a empirical semivariogram obtained from the data set for the $PM_{2.5}$ levels from 01 August 2015 to 15 March 2016. In particular, the cutoff (or separation) distance up to which point pairs are included in semivariance estimates is 1Km. The width of subsequent distance intervals into which data point pairs are grouped is 0.1Km and the time lags come from 0 to 5 days. The empirical semivariogram can be thought of as a measure of dissimilitude in space/time. The lower the semivariogram value, the higher the correlation between two pairs of data points. One can note that a temporal correlation up to 5 days apart is observed and a strong spatial correlation up to 0.6 Km is present. In addition, a clear nugget effect is observed due to the noisy data. Finally, one can conclude that the plot exhibits some spatial-temporal interaction.

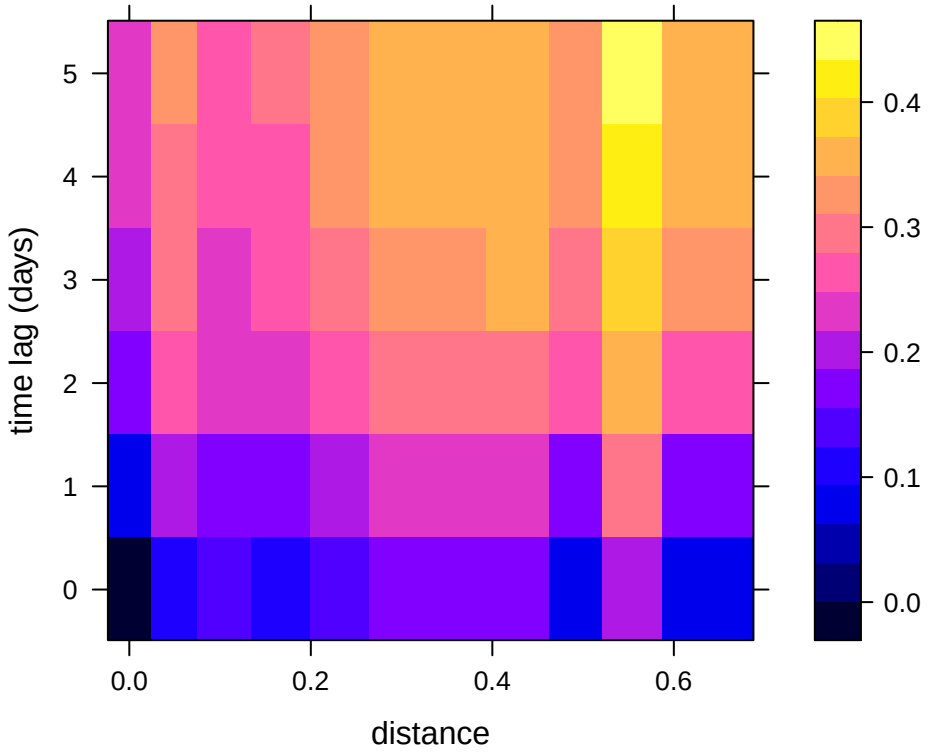

Figure 4-6: Empirical spatio-temporal semivariogram of daily $PM_{2.5}$ data set from 01 August 2015 to 15 March 2016. The plot is produced with cutoff=1Km, bin width=0.1Km and time lags from 0 to 5 days.

4.5 Discussion

In this chapter, a summary of methods used for spatio-temporal modelling and in particular spatial interpolation is presented. The concept of stationarity and separability are introduced, which is an assumption that can simplify the process of spatial (and temporal) prediction. However, the effects of making these assumptions need to be understood and decisions made on whether they are tenable. Often, there may not be enough evidence to fully justify such assumptions, but pragmatic choices may have to be made in order to achieve desired outcomes. In such cases, it is important that any modelling decisions are taken into account when drawing conclusions, and in using the outputs (e.g. spatial predictions) in further analyses (e.g. estimating risks in epidemiological studies). In addition to the description of the theoretical aspects of spatial and temporal prediction, analysis of data from the LAQN is presented using

empirical spatio-temporal covariograms and semi-variograms. The application of these methods to a selection of data (from August 2015 to March 2016) showed the presence of temporal correlation up to 5 days apart, and a strong spatial correlation up to 0.6 KM after which it reduces over further distance.

Chapter 5

Dynamic Spatio-Temporal Models

Different approaches have been proposed to model spatio-temporal processes. Traditional spatial statistics techniques consider time as an extra dimension, see Cressie (1993). But, these approaches do not consider the differences between space and time. From a multivariate geostatistical perspective, one should specify convenient space-time covariance functions. However, this approach is limited by the fact that such covariance functions are often not realistic for more complex models, see Cressie & Huang (1999). In section 6 two covariance-based models were proposed.

Spatio-temporal processes can also be considered from the multiple time series perspective, Lütkepohl (2005). That is, each spatial location is associated with a time series. Then, multivariate time series techniques can be expressed as a space-time problem. However, such approaches do not notice the differences between space and time, and prediction at locations for which data were not observed is limited. Moreover, such methods are difficult to implement in cases where the dimensionality of the number of spatial locations is high.

In this chapter, the focus is on hierarchical spatio-temporal models. These methods offer several advantages over traditional approaches. Firstly, physical and dynamical components can be easily incorporated into the conditional formulation. The consideration of simple and more realistic conditional models leads to a more complicated spatio-temporal covariance structure than can be specified directly. Moreover, by making use of the sparse structure inherent in the hierarchical approach, the models can be computed with larger data sets. This modelling approach can be helpful for a great va-

riety of problems in the air pollution. The idea behind dynamic spatio-temporal models (DSTM) is the hierarchical (HM) state-space perspective that has been described in chapter 3.

A natural approach to spatio-temporal modelling for complex dynamical processes is a combination of spatial and time series techniques. It can be accomplished by a spatio-temporal dynamic model formulation, see Cressie & Wikle (2011, 2019). However, the estimation can be problematic due to the high dimensionality of the state process in DLM context. One alternative is is to reduce dimensionality by projecting the state-process on some set of spectral basis functions, see Mardia *et al.* (1998); Wikle & Cressie (1999) for further details. Another option is specifying a simple random walk dynamics such as in Huerta *et al.* (2004). Alternatively, if one knows explicitly the model parameters, they can be incorporated directly into the parameterisation, Wikle *et al.* (2001).

The standard approach where model parameters are unknown uses the EM algorithm to estimate parameters, Shumway & Stoffer (1982). Estimation in the spatio-temporal dynamical model setting can be achieved through a state-space framework. Given parameters, the unobserved state-process can be estimated via the Kalman filter or Kalman smoother. In Wikle & Cressie (1999); Xu & Wikle (2007) estimation and prediction using Kalman filtering/smoothing and reduced-dimension processes are performed. An review of spatio-temporal Kalman filter implementations can be seen in Cressie & Wikle (2011).

In this chapter, an efficient estimation approach for spatio-temporal dynamic models is described, in which the covariance matrices are parameterised. The EM algorithm to carry out this estimation is implemented. In addition, the idea of spatio-temporal dimension reduction is introduced. The key is to change the state vector in a lower dimensional space by using a spectral basis. Although one could use any set of orthonormal basis functions such as Fourier, wavelets, or empirical functions, the empirical orthogonal functions (EOFs) are chosen for this chapter. This selection is convenient since they are widely used in meteorological studies, see chapter 5 in Cressie & Wikle (2011).

5.1 Hierarchical DSTM Models

In this section, a general scheme of hierarchical modelling in the context of a DSTM is given. If it recognised that observing a process perfectly is not possible, then it is required mapping that relates observations to the "true" process. In the context of

DSTMs, one could specify a *data model* level that gives a model for the data, conditioned on the true process of interest and some parameters. At the next level, a *process model* can specify the dynamic evolution of the (hidden/latent/state) process, given some parameters. At the third level, one can have *parameters models* for the parameters from the previous two stages (Bayesian hierarchical model, BHM), or estimates of the parameters (empirical hierarchical model, EHM).

5.1.1 Data Models

The distribution of the data can be written as

$$[\{Z(\mathbf{x};r):\mathbf{x}\in D_s, r\in D_t\}|\{Y(\mathbf{s};t):\mathbf{s}\in\mathcal{N}_x, t\in\mathcal{N}_r\},\boldsymbol{\theta}_D], \tag{5.1}$$

where $Z(\mathbf{s};r)$ is an observation at spatial location $\mathbf{s}$ and time r, $Y(\mathbf{s};t)$ represents the (hidden or latent) process of interest at spatial location $\mathbf{s}$ and time t, and $\boldsymbol{\theta}_D$ refers to data-model parameters than can vary in space and/or time. The *spatial neighbourhood* of $\mathbf{x}$ can be represented as $\mathcal{N}_x$ and, the *past neighbourhood* of r as $\mathcal{N}_r$. That is, it is the observations at possible different supports, conditional on values of the true process of interest at some other locations of support and some parameters. The form of the DSTM data model shown in 5.1 is extremely general to be useful. It is more convenient to give specific types of models that are more useful for spatio-temporal models and that show the flexibility of the conditional framework.

5.1.2 Process Models

The DSTM process distribution can be written as

$$[Y(\mathbf{s};t)|\{Y(\mathbf{w};t-\tau_1):\mathbf{w}\in\mathcal{N}_s^{(1)}\},\ldots,\{Y(\mathbf{w};t-\tau_p):\mathbf{w}\in\mathcal{N}_s^{(p)}\},\boldsymbol{\theta}_P], \tag{5.2}$$

where $\mathcal{N}_{\mathbf{s}}^{(1)},\ldots,\mathcal{N}_{\mathbf{s}}^{(p)}$ are neighbourhoods of spatial locations $\mathbf{s}$, corresponding to the lags $0<\tau_1<\ldots<\tau_p$, and $\boldsymbol{\theta}_P$ refers to the process-model parameters. Again, the form of 5.2 is general but not necessarily useful, in section 5.2 specific formulations will be discussed. One the most relevant problems of using DSTMs is the computational burden due to the dimensionality.

5.1.3 Parameter Models

Finally, the parameter-model distribution is presented as

$$[\boldsymbol{\theta}_D,\boldsymbol{\theta}_P|\boldsymbol{\theta}_h], \tag{5.3}$$

where $\boldsymbol{\theta}_h$ refers to hyperparameters. This is a general framework but some assumptions need to be made to have more useful models, as one can see in the next section. in many cases, it is assumed conditional independence in the parameters distributions, such as $[\boldsymbol{\theta}_D, \boldsymbol{\theta}_P|\boldsymbol{\theta}_h] = [\boldsymbol{\theta}_D, \boldsymbol{\theta}_h][\boldsymbol{\theta}_P|\boldsymbol{\theta}_h]$. One should notice that in some cases it would be sufficient to specify or estimate any of the three levels, rather than model them at this third level of hierarchy. Moreover, one could need even lower levels in more complicated scenarios. It is the case that the choice of these prior distributions is driven by the underlying specific problem.

5.2 Latent Linear Gaussian DSTMs

DSTM are widely used models in which the process models are assumed to have additive Gaussian error distributions and linearity in an evolution operator. Unless otherwise stated, it is assumed that the locations can have either point or areal support. Different supports for the prediction locations and data locations can be also assumed.

From section 5.1 one can consider, $D_t = \{0, 1, 2, \ldots\}$, as discrete time with temporal domain with a constant time increment. One can denote the data and potential data by $\{Z_t(\mathbf{s}) : \mathbf{s} \in D_\mathbf{s}; t = 0, 1, \ldots\}$; the latent process is denoted by $\{Y_t(\mathbf{s}) : \mathbf{s} \in D_\mathbf{s}; t = 0, 1, \ldots\}$, where $D_\mathbf{s}$ is a spatial (discrete/continuous) domain. About $Z_{t_0}(\mathbf{s}_0)$, there is no available data but even in this case one can make inference on $Y_{t_0}(\mathbf{s}_0)$.

5.2.1 Data Models

The relationship between the observations and the latent spatio-temporal process can be described through a *data model*,

$$Z_t(\cdot) = \mathcal{H}_t(Y_t(\cdot), \boldsymbol{\theta}_{D,t}, \epsilon_t(\cdot)), \qquad t = 1, \ldots, T, \tag{5.4}$$

where $Z(\cdot)$ corresponds to the data at time t, and $(\cdot)$ represents any spatial location. The mapping $\mathcal{H}_t$ (linear or not linear) connects the data to the latent process. The spatio-temporal variability, is given by $\epsilon_t(\cdot)$, and data-model parameters are represented by the vector, $\boldsymbol{\theta}_{D,t}$.

The power of the data model comes from conditioning, if one assume that $Z(\cdot)$ are independent in time. When conditioned on the true process, $Y(\cdot)$, parameters $\boldsymbol{\theta}_{D,t}$ have the joint distribution represented in product form,

$$[\{Z_t(\cdot)\}_{t=1}^T|\{Y_t(\cdot)\}_{t=1}^T, \{\boldsymbol{\theta}_{D,t}\}_{t=1}^T] = \prod_{t=1}^T [Z(\cdot)|Y(\cdot), \boldsymbol{\theta}_{D,t}]. \tag{5.5}$$

It will be considered that the component distributions on the right-hand side of the equation 5.5 to be Gaussian, but other members of the exponential family of distributions can also be considered.

5.2.2 Process Models

If a first-order *Markov assumption* is made for the *process model*, when conditioning on the past, only the recent past is important to explain the present. One have the next simplification,

$$[Y_t(\cdot)|Y_{t-1}(\cdot),\ldots,Y_0,\{\boldsymbol{\theta}_{P,t}\}_{t=0}^{T}] = [Y_t(\cdot)|Y_{t-1}(\cdot),\boldsymbol{\theta}_{P,t}], \qquad t=1,2,\ldots \tag{5.6}$$

If one uses the chain rule, then the joint distribution can be written as,

$$[Y_0(\cdot),Y_1(\cdot),\ldots,Y_n(\cdot)] = [Y_T(\cdot)|Y_{T-1}(\cdot),\ldots,Y_0(\cdot)]\ldots[Y_1(\cdot)|Y_0(\cdot)][Y_0(\cdot)], \tag{5.7}$$

under this assumption and considering equation 5.6, one have a simpler expression for the conditioned joint distribution

$$[Y_0(\cdot),Y_1(\cdot),\ldots,Y_T(\cdot)|\{\boldsymbol{\theta}_{P,t}\}_{t=0}^{T}] = \left(\prod_{t=1}^{T}[Y_t(\cdot)|Y_{t-1}(\cdot),\boldsymbol{\theta}_{P,t}]\right)[Y_0(\cdot)|\boldsymbol{\theta}_{P,0}]. \tag{5.8}$$

A stochastic model for $Y_t(\cdot)$ can be specified, in general, a *dynamical model* can be written as

$$Y(\cdot) = \mathcal{M}(Y_{t-1}(\cdot),\boldsymbol{\theta}_{P,t},\eta_t(\cdot)), \qquad t=1,2,\ldots, \tag{5.9}$$

where $\mathcal{M}$ is a function at the previous time, called *evolution operator*. Meanwhile, $\boldsymbol{\theta}_{P,t}$ are parameters (possibly with spatial or temporal dependence) that describe the dynamical transition and $\eta_t(\cdot)$ is a spatial error term process that is independent in time. One should notice that a distribution for the conditioned initial state, $Y_0(\cdot|\boldsymbol{\theta}_{P,0})$, needs to be specified or or conditioned on it. The evolution operator can be linear or nonlinear and the associated distribution can be Gaussian or non-Gaussian with additive or multiplicative errors. Higher-order Markov assumptions can be considered but this formulation could increase dimensionality. First-order representation in terms of probability distributions in general is enough.

5.2.3 Parameters Models

A Bayesian approach requires distributions to be assigned to the parameters defined in the data model and the process model, $\{\boldsymbol{\theta}_{D,t}, \boldsymbol{\theta}_{P,t}\}$. Spatially or temporally varying dependence on covariate information should be specified for the parameters. As mentioned in section 3, one of the most important aspects of Bayesian hierarchical modelling is the specification of these distributions. In that case, an important modelling challenge in DSTMs is to come up with ways to effectively reduce the parameter space.

In some cases, it is possible to estimate the parameters in an EHM context. If one decide to follow this approach, it is not needed to specify a distribution for the parameters defined in the data model and the process model. This is frequently done in state-space models in time series and sometimes the expectation-maximisation (EM) algorithm is implemented. Ir that the choice of the estimation approach is a specific problem.

5.3 Linear Data Model with Additive Gaussian Error

Let us suppose some data at locations $\{\mathbf{r}_{jt} : j = 1, \ldots, m_t; t = 0, 1, \ldots, T\}$. A different number of data locations for each observation time can be considered, with a finite set of m possible data locations, i.e., $m_t < m$. Let $\{\mathbf{s}_i : i = 1, \ldots, n\}$ a set of locations associated to a Gaussian latent process $Y_t(\mathbf{s}_i)$, and $\mathcal{M}$ the linear evolution operator, both defined as in the equation 5.9.

Consider the m_t-dimensional data vector, $\mathbf{Z}_t \equiv (Z(\mathbf{r}_{1t}), \ldots, Z(\mathbf{r}_{m_t t}))'$, and the n-dimensional latent process vector, $\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_n))'$. The linear data model with additive Gaussian error for the jth observation at time t, is written as

$$Z_t(\mathbf{r}_{jt}) = b_t(\mathbf{r}_{jt}) + \sum_{i=1}^{n} h_{t,ji} Y_t(\mathbf{s}_i) + \epsilon_t(\mathbf{r}_{jt}), \tag{5.10}$$

for $t = 1, \ldots, T$, where $b_t(\mathbf{r}_{jt})$ is an additive offset term for the jth observation at time t, $\{h_{t,ji}\}_{i=1}^{n} \equiv \mathbf{h}'_{t,j}$ are coefficients that map the latent process to the jth observation at time t, and the error term $\epsilon_t(\cdot)$ is independent of $Y_t(\cdot)$. Since $j = 1, \ldots, m_t$, the data model can be written in vector matrix form as

$$\mathbf{Z}_t = \mathbf{b}_t + \mathbf{H}_t \mathbf{Y}_t + \varepsilon_t, \qquad \varepsilon_t \sim iid\ \ Gau(\mathbf{0}, \mathbf{C}_{\epsilon,t}), \tag{5.11}$$

where $\mathbf{b}_t$ is the m_t-dimensional offset term, $\mathbf{H}_t$ is the $m_t \times n$ matrix mapping, which jth row corresponds to $\mathbf{h}'_{t,j}$, and $\mathbf{C}_{\epsilon,t}$ is an $m_t \times m_t$ error covariance matrix.

5.3.1 Latent Spatio-Temporal Dynamic Process

The main issue with dynamic spatio-temporal process, $\mathbf{Y}_t$, is specifying how the process transitions from one time to the next, this is where our modelling effort shall be focused. Unless specified otherwise, it is assumed $\mathbf{Y}_t$ has a mean of zero; and as before, a first-order Markov assumption can be made to describe its evolution. A further description of this component will be given after.

5.3.2 Additive Offset Term

Despite the simplicity in the general DSTM framework, there are circumstances where one could include another level of sophistication, such as an additive bias in the observations. If one is interested, it is possible to infer the additive offset term, $\mathbf{b}_t$, from 5.11, some assumptions can be made with this in mind. It is possible that the additive offset term, $b_t(\mathbf{r}_{jt})$, can be fixed through time, space, or constant across space and time; that is $b_t(\mathbf{r}_{jt}) \equiv b(\mathbf{r}_{jt})$, $b_t(\mathbf{r}_{jt}) \equiv b_t$ or $b_t(\mathbf{r}_{jt}) \equiv b$ respectively. It can also be defined it in terms of covariates $b_t(\mathbf{r}_{jt}) \equiv \mathbf{x}'_{t,j}\boldsymbol{\beta}$ or $b_t(\mathbf{r}_{jt}) \equiv \mathbf{x}'_t\boldsymbol{\beta}_j$ where $\mathbf{x}_{t,j}$ and $\mathbf{x}_t$ are q-dimensional vectors of covariates and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_j$ are q-dimensional parameter vectors. One can also consider the offset parameters to be either spatial or temporal random processes with distributions assigned at another level of hierarchy into the model. In this case, $\mathbf{b}_t \sim Gau(\mathbf{X}_t\boldsymbol{\beta}_t, \mathbf{C}_b)$ where $\mathbf{C}_b$ is a positive-definite matrix such as it was described in chapter 6.

5.3.3 Observation Mapping Matrix/Function

The observation mapping matrix, $\mathbf{H}_t$, is varying in time and can accommodate potential observation networks at each time. One of the simplest cases is when it is an *incidence matrix* of zeros and ones. In this case, data location is associated with one or more of the process locations. The incidence matrix can accommodate missing data or can serve as an interpolator such that each observation is related to some weighted combination of the process values. It is possible to parameterise the mapping matrix and/or estimate it directly in some cases, but it is more common to assume that it is known.

There is another situation where one can know the elements of the incidence vector, $\mathbf{h}_{t,j}$, it is called the *change-of-support* problem. In situations where this change is used, implications for the measurement covariance matrix should be taken into account and it is also possible that the weights $\mathbf{H}_t$ will change with the time. See chapter 7 of Cressie & Wikle (2011), for details.

5.3.4 Error Covariance Matrix

The additive error process $\{\epsilon_t(\mathbf{r}_{jt})\}$ is assumed to have mean of zero, is Gaussian, and usually one assumes that the errors are independent in time, but they could include dependence in space or time. Let $\varepsilon_t \equiv (\epsilon_t(\mathbf{r}_{1t})), \ldots, \epsilon_t(\mathbf{r}_{mt})$, where $\epsilon(\cdot)$ is defined at a finite set of m_t observation locations. Now our goal shall be to specify time-varying covariance matrices $\{\mathbf{C}_{\epsilon,t}\}$, in practice, one can assume that these data-model errors are time/space independent with constant variance, i.e., $\mathbf{C}_{\epsilon,t} = \sigma^2_\epsilon \mathbf{I}_{m_t}$, where σ^2_ϵ is the measurement-error variance. Stating a covariance matrix depends on the specific problem, but one can take advantage of hierarchical modelling to simplify the problem.

There are some scenarios where the data-model error covariance matrix is assumed constant over time, $\mathbf{C}_{\epsilon,t} = \mathbf{C}_\epsilon$, and $m_t = m$, in this case one could try to estimate $\mathbf{C}_\epsilon$ or parameterise it in terms of some valid spatial covariance functions if if there is spatial dependence.

5.4 Process Model

Specifying a transition process in a spatio-temporal statistical setting can be difficult due to the inconvenience of high dimensionality. The linear-evolution equation with Gaussian errors describes how the process at location j at the previous time influences the process at location i at the current time, it can be written in a matrix form,

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta_t}, \qquad \boldsymbol{\eta}_t \sim iid\ \ Gau(\mathbf{0}, \mathbf{C}_\eta). \tag{5.12}$$

This is a particular version of the process shown in 5.9, where the $n \times n$ transition matrix is given by $\mathbf{M}$ with (possible unknown) elements $\{m_{ij}\}$ and the additive spatial error process, $\{\boldsymbol{\eta}_t\}$, independent over time and Gaussian with mean zero and covariance matrix $\mathbf{C}_\eta$; $\boldsymbol{\eta}_t$ is independent of $\mathbf{Y}_{t-1}$. Unless specified otherwise, one shall assume $\{\mathbf{Y}_t\}$ has a mean of zero throughout this work. The stability condition in this case requires that the maximum modulus of the eigenvalues of $\mathbf{M}$ be less than 1, see appendix A.

Consider the ith element of $\mathbf{Y}_t$ and the associated evolution implied by the process 5.12,

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^{n} m_{ij} Y_{t-1}(\mathbf{s}_j) + \eta(\mathbf{s}_i), \tag{5.13}$$

where the weights m_{ij} correspond to the ith row and the jth column of $\mathbf{M}$. It is unlikely having a real world problem where one knows the elements of $\mathbf{M}$, but when n is relatively small, one can estimate $\mathbf{M}$. In spatio-temporal dynamic modelling, the

fundamental problem is how $\mathbf{M}$ can be parameterised with relatively few parameters. In general, obtaining reliable estimates of the $n \times n$ propagator matrix $\mathbf{M}$ can be challenging. However, $\mathbf{M}$ and $\mathbf{C}_\eta$ are often assumed to depend on parameters $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_\eta$ respectively, to reduce the curse of dimensionality. In some applications of fairly low dimensionality and large sample sizes (when n is small and $T \gg n$) one can estimate $n \times n$ matrices $\mathbf{M}$ and $\mathbf{C}_\eta$ directly in an EHM. This is commonly done in state-space models of time series, in particular, this approach will be followed.

There are other approaches that one can consider to parameterise the evolution matrix for the first-order, some of these are motivated by the underlying dynamical process. See chapter 7.2 Cressie & Wikle (2011), for details. Some of these approaches are briefly described here. In particular, the approach that will be followed later in section 5.9 is the integro-difference equation (IDE).

5.4.1 Spatio-Temporal Random Walk

Perhaps one of the simplest parameterisation of DSTM process. It is a multivariate extension of the random walk. That is,

$$\mathbf{Y}_t = \mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim iid \;\; Gau(\mathbf{0}, \mathbf{C}_\eta). \tag{5.14}$$

In this case, the process value at location $\mathbf{s}_i$ and time t is the value at the same location $\mathbf{s}_i$ at the previous time $(t-1)$, plus some noise. It is clear that $\mathbf{M} = \mathbf{I}$ in the equation 5.12. There is a possibility that term $\boldsymbol{\eta}_t$ is spatially dependent. The advantage of this parameterisation is that there are few parameters. But the disadvantage is that it is not stationary in time. Then, this parameterisation is not realistic for most of the process on large spatial and temporal scales.

5.4.2 Spatial Autoregresive Model

A natural extension of the simple autoregressive model is to allow the AR parameters to vary heterogeneously across space. For example, for n locations it can be written as,

$$\mathbf{Y}_t = \text{diag}(\mathbf{m})\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim iid \;\; Gau(\mathbf{0}, \mathbf{C}_\eta), \tag{5.15}$$

where $\mathbf{m} \equiv (m(\mathbf{s}_1), \ldots, m(\mathbf{s}_n))'$. Once more, the advantage is that there are relatively few parameters and it is fairly to estimate the parameters. But this model can not handle direct interactions between space and time for multiple spatial locations.

5.4.3 "Lagged Nearest-Neighbour" Model

Let $\mathbf{M}$ be parameterised such that only the m_{ij} corresponding to the location $\mathbf{s}_i$ and its nearest neighbour $\{\mathbf{s}_j : j \in \mathcal{N}_i\}$ at the previous time are important. This assumption can be made in the context of a process with areal support or on a regular grid, with point or areal support. The model can be written as

$$Y_t(\mathbf{s}_i) = \sum_{j \in \mathcal{N}_i} m_{ij} Y_{t-1}(\mathbf{s}_j) + \eta_t(\mathbf{s}_i), \tag{5.16}$$

where $\mathcal{N}_i$ refers to a specified "neighbourhood" of $\mathbf{s}_i, 1, \ldots, n$. In general, it can be assumed that errors are Gaussian: $\boldsymbol{\eta}_t \sim iid\ Gau(\mathbf{0}, \mathbf{C}_\eta)$. The matrix $\mathbf{M}$ is sparse, many of its elements are zero. Even in this situation, the estimation of the nonzero m_{ij} can be difficult.

5.4.4 PDE-Based Parameterisation

For some physical and environmental processes, the underlying spatio-temporal process can follow partial differential equations (PDEs) or ordinary differential equations (ODEs). In the case of linear PSEs, standard finite differencing implies linear Markov spatio-temporal models. Let us consider the general PDE,

$$\frac{\partial Y}{\partial t} = \mathcal{M}(Y, \omega, \boldsymbol{\theta}), \tag{5.17}$$

where $\mathcal{M}$ is some function of the variable of interest, Y, other potential variables, ω, and process model parameters, θ. It is a deterministic model.

5.4.5 Integro-difference equation (IDE)

In the context of linear Markovian spatio-temporal process models, the value of the process at a given location at the present time is determined by the past. It is made up of a weighted combination of the process throughout the spatial domain at previous times, plus an additive Gaussian innovation. The process can be represented in a continuous-spatial context through an *integro-difference equation* (IDE). A first-order spatio-temporal IDE process model is given by

$$Y_t(\mathbf{s}) = \int_{D_\mathbf{s}} m(\mathbf{s}, \mathbf{x}, \boldsymbol{\theta}_p) Y_{t-1}(\mathbf{x}) \mathrm{d}\mathbf{x} + \eta_t(\mathbf{s}) \quad \mathbf{s}, \mathbf{x} \in D_\mathbf{s} \tag{5.18}$$

for $t = 1, 2, \ldots$. In this case, $m(\mathbf{s}, \mathbf{x}, \boldsymbol{\theta}_p)$ is a *transition kernel* depending on parameters $\boldsymbol{\theta}_p$ that specify weights for the process at the previous time over the spatial domain. $D_\mathbf{s}$

and $\eta_t(\cdot)$ is a time-varying continuous mean-zero Gaussian spatial process independent of $Y_{t-1}(\cdot)$.

If it is supposed a finite set of prediction spatial locations (or regions) $D_{\mathbf{s}} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. The first-order IDE evolution process model 5.18 can be discretised and written as a stochastic difference equation,

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^{n} m_{ij}(\boldsymbol{\theta}_p) Y_{t-1}(\mathbf{s}_j) + \eta(\mathbf{s}_i) \tag{5.19}$$

for $t = 1, 2, \ldots$. And with transition weights $m_{ij}(\boldsymbol{\theta}_p)$ that depend on parameters $\boldsymbol{\theta}_p$. In particular, the process at $Y(\mathbf{s}_i)$ considers a weighted combination of the values of the process at time $t-1$ and at a discrete set of spatial locations.

The equation 5.19 suggested potential parameterisations of $\mathbf{M}$ in terms of few parameters. Some of these parametrisations include nonzero off-diagonal elements. Usually, the propagator matrix is not symmetric, $m_{ij}(\boldsymbol{\theta}) \neq m_{ji}(\boldsymbol{\theta})$. Some disadvantages can be observed under this approach. For instance, even if $\boldsymbol{\theta}$ is known, the dimensionality of $\mathbf{M}$ can be computationally problematic. An alternative implementation can be more efficient. This can be done considering the IDE model in spectral form.

Now, if the process vector is denoted as $\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), Y_t(\mathbf{s}_2), \ldots, Y_t(\mathbf{s}_n))'$, then the equation 5.19 can be written in vector–matrix form as a linear first-order vector autoregression DSTM,

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t. \tag{5.20}$$

The $n \times n$ transition matrix is given by $\mathbf{M}$ with elements $\{m_{ij}\}$, and the additive spatial error process $\boldsymbol{\eta}_t \equiv (\eta_t(\mathbf{s}_1), \eta_t(\mathbf{s}_2) \ldots, \eta_t(\mathbf{s}_n))$ is independent of $\mathbf{Y}_{t-1}$ and is specified to be mean-zero and Gaussian with spatial covariance matrix $\mathbf{C}_\eta$.

Then, $\mathbf{Y}_t$ can be expanded in terms of some complete class of spatial spectral basis vectors, given by the columns of $\boldsymbol{\Phi}$, one can obtain,

$$\mathbf{Y}_t = \boldsymbol{\Phi}\boldsymbol{\alpha}_t,$$

where $\boldsymbol{\alpha}_t$ are the associated spectral coefficients. Note that it can be considered as a linear transformation of the spatial process at time t into the spectral domain. This

suggest the dynamical relationship

$$\boldsymbol{\alpha}_t = \mathbf{M}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\gamma}_t, \tag{5.21}$$

where $\mathbf{M}_\alpha \equiv (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{M}\boldsymbol{\Phi}$ is the spectral propagator matrix, and $\boldsymbol{\gamma}_t = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{\eta}_t$ is the spectrally transformed noise-forcing process. Note that if the basis vector are orthogonal, then 5.21 can be simplified considering $\boldsymbol{\Phi}$ such as $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \mathbf{I}$. It can represent some advantages since $\mathbf{M}_\alpha$ and the covariance for $\boldsymbol{\gamma}_t$ are often simpler than their equivalent versions in the spatial domain. See Wikle & Cressie (1999).

5.5 Multivariate DSTM

Most real-world processes are not isolated from other processes. For example, a model to analyse the behaviour of pollutants such as nitrogen dioxide (NO_2) and particle matter (PM), would be possibly better if their relationship is considered. Modelling joint statistical dependence could extend both joint and conditional formulations for multivariate DSTM process models.

5.5.1 Multivariate DSTMs: Augmenting the State Process

In dynamical perspective, all the approaches for modelling DSTMs are applicable in multivariate context. If one has a system of several processes in which first-order linear dynamics were appropriate, one could augment the state variable by concatenating the state vectors. Let us consider the process $\mathbf{Y}_t^{(k)}$ of dimension n_k; $k = 1, \ldots, K$. The augmented state is $\mathbf{Y}_t = (\mathbf{Y}_t^{(1)}, \ldots, \mathbf{Y}_t^{(K)})'$, one could use a dynamical model to capture its temporal variability,

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t; \qquad \boldsymbol{\eta}_t \sim iid\ \ Gau(\mathbf{0}, \mathbf{Q}_\eta), \tag{5.22}$$

which is of dimension $n_+ \equiv \sum_k n_k$. The propagator matrix $\mathbf{M}$ allows each process at the previous time to have influence on itself and the others at the current time. The error covariance matrix $\mathbf{Q}_\eta$ is in block form, corresponding to the covariance matrix of each process (the diagonal blocks) and the cross-covariance matrices (the off-diagonal blocks). Although dimensionality can be a potential problem, first-order lag structure and non-time-varying propagator matrices are sufficient for general multivariate modelling issues.

5.6 Parameter and process dimension reduction in dynamic spatio-temporal models

The linear Gaussian DSTM described in section 5.3 has unknown parameters associated with the data mode $\mathbf{C}_\eta$, the transition $\mathbf{M}$ and the initial-condition distribution ($\boldsymbol{\mu}_0$ and $\mathbf{C}_0$ for instance). For the linear Gaussian data model, a simple characterisation of $\mathbf{C}_\epsilon$ can be made, $\mathbf{C}_\epsilon = \sigma_\epsilon^2\mathbf{I}$, for example. Alternatively, the covariance matrix implied by a spatial random process can have just a few parameters such as a Mátern spatial covariance function or a spatial conditional autoregressive process. In a complex DSTM modelling context for the process-error spatial variance–covariance matrix, $\mathbf{C}_\eta$. One could use a spatial covariance-function representation such as a Matérn function or a basis-function random effects representation. One of the challenges when considering DSTMs in hierarchical statistical settings is the dimension associated with the process-model level of the DSTM. One way to proceed, is reducing the number of free parameters to be inferred in the model and/or reduce the dimension of the spatio-temporal dynamic process.

5.6.1 Parameter Dimension Reduction

The linear Gaussian DSTM described in section 5.3 has unknown parameters associated with the data mode $\mathbf{C}_\eta$, the transition $\mathbf{M}$ and the initial-condition distribution ($\boldsymbol{\mu}_0$ and $\mathbf{C}_0$ for instance). For the linear Gaussian data model, a simple characterisation of $\mathbf{C}_\epsilon$ can be made, $\mathbf{C}_\epsilon = \sigma_\epsilon^2\mathbf{I}$, for example. Alternatively, the covariance matrix implied by a spatial random process can have just a few parameters such as a Mátern spatial covariance function or a spatial conditional autoregressive process. In a complex DSTM modelling context for the process-error spatial variance–covariance matrix, $\mathbf{C}_\eta$, one could use a common spatial covariance-function representations such as a Matérn function or a basis-function random effects representation.

The transition-matrix parameters require particular attention, since there could be as many as n^2 of them. In simple linear cases of DSTM, the transition matrix $\mathbf{M}$, could be parameterised as a random walk ($\mathbf{M} = \mathbf{I}$), a spatially homogeneous autoregressive process ($\mathbf{M} = \theta_p\mathbf{I}$), or a spatially varying autoregressive process ($\mathbf{M} = \text{diag}(\boldsymbol{\theta}_p)$).

The last parameterisation can be more useful, for example, one could have the process model where $\mathbf{C}_\eta = \sigma_\eta^2\mathbf{I}$, and $\mathbf{M} = \text{diag}(\boldsymbol{\theta}_p)$. The conditional distribution can be

decomposed as,

$$[\mathbf{Y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}_p, \sigma_\eta^2] = \prod_{i=1}^{n}[Y_t(\mathbf{s}_i)|Y_{t-1}(\mathbf{s}_i), \theta_p(i), \sigma_\eta^2], \qquad t = 1, 2, \ldots \tag{5.23}$$

where $\boldsymbol{\theta_p} = (\theta_p(1), \ldots, \theta_p(n))'$. At one location, the transition is influenced just by the Y-value at the previous time, i.e., one has a spatially independent univariate AR(1) processes at each spatial location.

5.6.2 Dimension Reduction in the Process Model

For DSTM, process models could be helpful to consider a decomposition in terms of fixed effects and random effects in a basis-function expansion,

$$Y_t(\mathbf{s}) = \mathbf{x}_t(\mathbf{s})'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha}\phi_i(\mathbf{s})\alpha_{i,t} + \sum_{j=1}^{n_\xi}\psi_j(\mathbf{s})\xi_{j,t} + \nu(\mathbf{s}), \tag{5.24}$$

where $\mathbf{x}_t(\mathbf{s})'\boldsymbol{\beta}$ is a term with covariates and fixed components $\boldsymbol{\beta}$, the first basis-expansion term, $\sum_{i=1}^{n_\alpha}\phi_i(\mathbf{s})\alpha_{i,t}$, contains known spatial basis functions, $\{\phi_i(\cdot)\}$, and random effects, $\{\alpha_{i,t}\}$; the residual basis-expansion term, $\sum_{j=1}^{n_\xi}\psi_j(\mathbf{s})\xi_{j,t}$, where the basis functions, $\{\psi_j(\cdot)\}$, are assumed known, and the random effects, $\{\xi_{j,t}\}$, can be non-dynamic or can contain simple temporal behaviour. This component can be used to capture the variability associated with spatio-temporal process on different spatial scales. Finally, $\nu_t(\cdot)$ is assumed to be a Gaussian process with mean zero and independent in time. The model 5.24 provides a flexible decomposition of spatio-temporal processes that is focused on the random effects, $\{\alpha_{i,t}\}$.

Reductions in process dimension of the n-dimensional vector, $\mathbf{Y}_t$, can be helpful since one could get a lower-dimensional process, $\{\boldsymbol{\alpha}_t\}$, of dimension n_α where $n_\alpha \ll n$. It can be convenient to rewrite 5.24 in vector form,

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\Psi}\boldsymbol{\xi}_t, \tag{5.25}$$

where $\mathbf{X}_t$ is an $n \times p$ matrix that can be interpreted as a spatial offset and/or covariate effects, $\boldsymbol{\Phi}$, is an $n \times n_\alpha$ matrix of basis vectors related to the process, $\{\boldsymbol{\alpha}_t\}$, and $\boldsymbol{\Psi}$ is an $n \times n_\xi$ matrix of basis vectors related to the latent coefficient process $\{\boldsymbol{\xi}_t\}$. The vector $\{\boldsymbol{\xi}_t\}$ can have different dynamic characteristics than $\{\boldsymbol{\alpha}_t\}$, the error process $\{\boldsymbol{\nu}_t\}$ is Gaussian and assumed to have mean zero and it is usually independent.

The evolution of the latent process can proceed according to the linear equations in-

volving a transition matrix,

$$\boldsymbol{\alpha_t} = \mathbf{M}_\alpha \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim iid \;\; Gau(\mathbf{0}, \mathbf{C}_\eta) \tag{5.26}$$

where $\mathbf{M}_\alpha$ is the $n_\alpha \times n_\alpha$ transition matrix, and $\boldsymbol{\eta}_t$ are independent of $\boldsymbol{\alpha}_{t-1}$ and independent in time. The matrices $\mathbf{M}_\alpha$ and $\mathbf{C}_\eta$ in 5.26 usually have simple structure depending on the process. It should be noticed that even in this low-dimensional context, in many cases parameter-space reduction may still be necessary.

One could consider the simple structures that were discussed in 5.4 in the context of linear DSTM process. Random walks, independent AR models, nearest-neighbour models are examples of these structures. The parameterisation of these matrices (in particular $\mathbf{M}$) is one of the greatest challenges in DSTMs. It can be facilitated by using parameter models in a BHM. But, in other applications of lower dimensionality and large sample sizes, one can estimate $n \times n$ matrices $\mathbf{M}$ and $\mathbf{C}_\eta$ directly in an EHM. This procedure is frequently done in state-space models of time series. In particular, this approach will be followed in the last example in this chapter, for more details see Cressie & Wikle (2011).

It is also important that the transition operator be non-normal, i.e., $\mathbf{M}_\alpha{}'\mathbf{M}_\alpha \neq \mathbf{M}_\alpha\mathbf{M}_\alpha{}'$. In which case, the eigenvectors of $\mathbf{M}$ are non-orthogonal, but still stable. Nondiagonal transition matrices can be considered in most cases. If the basis functions given in $\boldsymbol{\Phi}$ are such that the elements of $\boldsymbol{\alpha}_t$ not spatially indexed such as some type of empirical orthogonal functions (EOFs). One can note that mechanistic knowledge can also be used in this case to motivate parameterisations for $\mathbf{M}_\alpha$. In the section 5.8, $\mathbf{M}_\alpha$ and $\mathbf{C}_\eta$ are estimated by the method of moments and by an EM algorithm.

5.7 Empirical Orthogonal Function (EOF) Analysis

Empirical orthogonal functions (EOFs) are geophysicist's terminology for the eigenvalue/eigenvector decomposition of covariance (or correlation) matrix (see Cressie & Wikle (2011) chapter 5). In statistics, EOF analysis is the spatio-temporal manifestation of principal component analysis (PCA). Depending on the application, EOFs can have different purposes:

1. In a diagnostic mode, to find principal spatial structures in terms of variance, along with the corresponding time variation of those structures.

2. To reduce dimensionality in large spatial-temporal data sets while simultaneously reducing noise.

One advantage of the EOF approach is that the complicated variability of spatio-temporal data can be compressed into a smaller set of eigenvectors.

EOFs can be calculated from the spectral decomposition of the empirical lag-0 spatial covariance matrix. However, they are more often obtained through a singular value decomposition (SVD), see appendix A and section 5.3.4 in Cressie & Wikle (2011).

For a spatio-temporal data set that does not have missing values, one can define $\mathbf{Z} \equiv [\mathbf{Z}_1, \ldots, \mathbf{Z}_T]'$, the $m \times T$ space-wide data matrix of data and let

$$\tilde{\mathbf{Z}} \equiv \frac{1}{\sqrt{T-1}}(\mathbf{Z} - \mathbf{1}_T \hat{\boldsymbol{\mu}}'_{Z,\mathbf{s}}), \tag{5.27}$$

the "detrended" and scaled data matrix, where $\mathbf{1}_T$ is a T-dimensional vector of ones and $\hat{\boldsymbol{\mu}}_{Z,\mathbf{s}}$ is the spatial mean vector given by 4.22. Given $\hat{\mathbf{C}}_Z^{(0)}$ is symmetric and nonnegative definite, the spectral decomposition,

$$\mathbf{C}_Z^{(0)} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}' = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}', \tag{5.28}$$

where $\boldsymbol{\Psi} \equiv (\psi_1, \ldots, \psi_m)$ is a matrix of spatially indexed eigenvectors given by the vectors $\boldsymbol{\psi}_k \equiv (\psi_k(\mathbf{s}_1), \ldots, \psi_k(\mathbf{s}_1))'$ for $k = 1, \ldots, m$ and $\boldsymbol{\Lambda} \equiv \text{diag}(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix of corresponding non-negative eigenvalues. Projections of the detrended data matrix onto the EOF basis functions, $\boldsymbol{\Psi}$, are given by

$$\mathbf{A} = (\sqrt{T-1})\tilde{\mathbf{Z}}\boldsymbol{\Psi}, \tag{5.29}$$

which are called the principal component (PC) time series. Their normalised version is given by $\mathbf{A}_{\text{norm}} \equiv \mathbf{A}\boldsymbol{\Lambda}^{-1/2}$, this normalisation allows the m time series to be plotted on the same scale, leaving their relative importance to be captured by their corresponding eigenvalues. The SVD calculation approach has some advantages:

- One does not need to calculate the empirical spatial covariance matrix.
- One gets the normalised PC time series and EOFs simultaneously.
- The procedure still works when $T < m$, although it can be problematic since $\mathbf{C}_Z^{(0)}$ is not positive-definite.

5.8 Implementation and Inference

If there is no prior knowledge on how to parameterise components in a hierarchical model, they can be estimated in the context of a standard state-space modelling framework. When taking this approach, it is important that a low dimensional representation of the spatio-temporal process is adopted. Since the dimension of the parameter space increases with the dimension of the process, therefore the model can become over parameterised.

Although it can be difficult to get analytical formulations for filtering, smoothing and forecast; in the case of a linear observation operator and Gaussian error distributions, one can get explicit distributions.

5.8.1 Kalman Filter

The *Kalman filter* is convenient for a sequential updating when one has linear operators and Gaussian error distributions, see West & Harrison (1997).

Let us define $\mathbf{Z}_{1:t} \equiv \{\mathbf{Z}_1, \ldots, \mathbf{Z}_t\}$ and $\mathbf{Y}_{0:t} \equiv \{\mathbf{Y}_0, \ldots, \mathbf{Y}_t\}$, for $t = 0, 1, \ldots, T$. So that, the conditional expectations for the filtering and forecast distributions can be defined as $\mathbf{Y}_{t|t} \equiv E[\mathbf{Y}_t|\mathbf{Z}_{1:t}]$ and $\mathbf{Y}_{t|t-1} \equiv E[\mathbf{Y}_t|\mathbf{Z}_{1:t-1}]$, respectively. The conditional error covariance matrices can be defined for filtering and forecasting as

$$\mathbf{P}_{t|t} \equiv E[(\mathbf{Y}_t - \mathbf{Y}_{t|t})(\mathbf{Y}_t - \mathbf{Y}_{t|t})'|\mathbf{Z}_{1:t}] \tag{5.30}$$

and

$$\mathbf{P}_{t|t-1} \equiv E[(\mathbf{Y}_t - \mathbf{Y}_{t|t-1})(\mathbf{Y}_t - \mathbf{Y}_{t|t-1})'|\mathbf{Z}_{1:t-1}]. \tag{5.31}$$

Consider the measurement distributions, $[\mathbf{Z}_t|\mathbf{Y}_t]$, given by

$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t; \qquad \boldsymbol{\varepsilon}_t \sim ind.\ Gau(\mathbf{0}, \mathbf{C}_{\varepsilon,t}), \tag{5.32}$$

where $\mathbf{H}_t$ is the observation operator that maps the process to the observations, and $\mathbf{C}_{\varepsilon,t}$ is the error covariance matrix, where the measurement errors are assumed to be independent across time.

The process distribution, $[\mathbf{Y}_t|\mathbf{Y}_{t-1}]$, can be considered and is given by the *vector autoregressive* model of order one, (VAR(1)):

$$\mathbf{Y}_t = \mathbf{M}_t\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t; \qquad \boldsymbol{\eta}_t \sim ind.\ Gau(\mathbf{0}, \mathbf{C}_{\eta,t}), \tag{5.33}$$

where $\mathbf{M}_t$ is the linear model operator that maps the evolution of the process in time, and $\mathbf{C}_{\eta,t}$ is a covariance matrix representing temporally independent stochastic features. In general, the measurement error process and the evolution error process are mutually independent for all time.

Using standard results for conditional-expectation and conditional-variance, the *forecast distribution* can be obtained by

$$\mathbf{Y}_t|\mathbf{Z}_{1:t-1} \sim Gau(\mathbf{Y}_{t|t-1}, \mathbf{P}_{t|t-1}), \tag{5.34}$$

where the mean vector and covariance are given by

$$\mathbf{Y}_{t|t-1} = E(E(\mathbf{Y}_t|\mathbf{Y}_{t-1})|\mathbf{Z}_{1:t-1}) = E(\mathbf{M}_t\mathbf{Y}_{t-1}|\mathbf{Z}_{1:t-1}) = \mathbf{M}_t\mathbf{Y}_{t-1|t-1}, \tag{5.35}$$

and

$$\begin{aligned}\mathbf{P}_{t|t-1} &= E(\text{var}(\mathbf{Y}_t|\mathbf{Y}_{t-1})|\mathbf{Z}_{1:t-1}) + \text{var}(E(\mathbf{Y}_t|\mathbf{Y}_{t-1})|\mathbf{Z}_{1:t-1}) \\ &= E(\mathbf{C}_{\eta,t}|\mathbf{Z}_{1:t-1}) + \text{var}(\mathbf{M}_t\mathbf{Y}_{t-1}|\mathbf{Z}_{1:t-1}) \\ &= \mathbf{C}_{\eta,t} + \mathbf{M}_t\mathbf{P}_{t-1|t-1}\mathbf{M}_t'.\end{aligned} \tag{5.36}$$

The *filtering distribution* is given by

$$\mathbf{Y}_t|\mathbf{Z}_{1:t} \sim Gau(\mathbf{Y}_{t|t}, \mathbf{P}_{t|t}), \tag{5.37}$$

where

$$\mathbf{Y}_{t|t} = \mathbf{P}_{t|t}(\mathbf{H}_t'\mathbf{C}_{\varepsilon,t}^{-1}\mathbf{Z}_t + \mathbf{P}_{t|t}^{-1}\mathbf{Y}_{t|t-1}) \tag{5.38}$$

and

$$\mathbf{P}_{t|t} = (\mathbf{H}_t'\mathbf{C}_{\varepsilon,t}^{-1}\mathbf{H}_t + \mathbf{P}_{t|t-1}^{-1})^{-1}. \tag{5.39}$$

The mean and variance of 5.37 can be written as

$$\mathbf{Y}_{t|t} = \mathbf{Y}_{t|t-1} + \mathbf{K}_t(\mathbf{Z}_t - \mathbf{H}_t\mathbf{Y}_{t|t-1}) \tag{5.40}$$

and

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t|t-1}, \tag{5.41}$$

where $\mathbf{K}_t$ is called the *Kalman gain* and is given by

$$\mathbf{K}_t \equiv \mathbf{P}_{t|t-1}\mathbf{H}_t'(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \mathbf{C}_{\varepsilon,t})^{-1}. \tag{5.42}$$

See Cressie & Wikle (2011), p. 446, for details.

If one assumes that the matrices $\mathbf{H}_t$, $\mathbf{M}_t$, $\mathbf{C}_{\varepsilon,t}$ and $\mathbf{C}_{\eta,t}$ are known, for $t = 0, 1, \ldots, T$ with initial conditions $\mathbf{Y}_{0|0} \equiv \mu_0$, $\mathbf{P}_{0|0} \equiv \mathbf{\Sigma}_0$, the *Kalman filter algorithm* can be used to obtain sequential estimates of states.

Algorithm: Kalman Filter
Set initial conditions: $\mathbf{Y}_{0|0} = \boldsymbol{\mu}_0$ and $\mathbf{P}_{0|0} = \mathbf{C}_{\eta,0}$
for $t = 1$ to T **do**
1. Forecast distribution step:
   (a) Obtain $\mathbf{Y}_{t|t-1} = \mathbf{M}_t \mathbf{Y}_{t-1|t-1}$
   (b) Obtain $\mathbf{P}_{t|t-1} = \mathbf{C}_{\eta,t} + \mathbf{M}_t \mathbf{P}_{t-1|t-1} \mathbf{M}_t'$
2. Filtering distribution step:
   (a) Obtain the Kalman gain, $\mathbf{K}_t \equiv \mathbf{P}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t' + \mathbf{C}_{\varepsilon,t})^{-1}$
   (b) Obtain $\mathbf{Y}_{t|t} = \mathbf{Y}_{t|t-1} + \mathbf{K}_t (\mathbf{Z}_t - \mathbf{H}_t \mathbf{Y}_{t|t-1})$
   (c) Obtain $\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_{t|t-1}$
end for

5.8.2 Kalman Smoother

One could be interested in the distribution of $\mathbf{Y}_t$ at time $t < T$, given all data up to T, this is the *smoothing distribution*, $[\mathbf{Y}_t | \mathbf{Z}_{1:T}]$. It is useful for retrospective analysis, for this purpose one can use the *Kalman smoother* that can be derived from various perspectives. See Cressie & Wikle (2011), p. 449 for details.

This smoothing distribution is denoted by

$$\mathbf{Y}_t | \mathbf{Z}_{1:T} \sim Gau(\mathbf{Y}_{t|T}, \mathbf{P}_{t|T}), \tag{5.43}$$

and, if one saves the results from the Kalman filter, the *Kalman smoother algorithm* is given by,

Algorithm: Kalman Smoother

Obtain: $\mathbf{Y}_{t|y}$, $\mathbf{P}_{t|t}$, $\mathbf{Y}_{t+1|t}$ and $\mathbf{P}_{t+1|t}$, for $t = 0, \ldots, T$ from the Kalman filter algorithm

for $t = T - 1$ to 0 **do**

1. Obtain $\mathbf{J}_t \equiv \mathbf{P}_{t|t}\mathbf{M}'_{t+1}\mathbf{P}^{-1}_{t+1|t}$
2. Obtain $\mathbf{Y}_{t|T} = \mathbf{Y}_{t|t} + \mathbf{J}_t(\mathbf{Y}_{t+1|T} - \mathbf{Y}_{t+1|t})$
3. Obtain $\mathbf{P}_{t|T} = \mathbf{P}_{t|t} + \mathbf{J}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{J}'_t$

end for

5.9 Estimation and prediction for Dynamic Spatio-Temporal Models

In some engineering applications, all parameters matrices are known but this is almost never the case for other contexts. In the case where the parameters associated with the evolution distribution are unknown, one can estimate them. This can be made either through empirical hierarchical modelling (EHM) such as in Wikle & Cressie (1999); Kang & Cressie (2011). Bayesian hierarchical modelling (BHM) could be used, the parameters are assumed to be random are given prior distributions and the estimation issues could be done via the posterior distributions, Berliner *et al.* (2000). Based on section 5.5, the proposed models are considering not just $PM_{2.5}$ concentrations but also PM_{10} and NO_2 data. This extension is relatively easy to implement.

If one has no prior knowledge about how to parameterise $\mathbf{M}$, then $\mathbf{M}$ can be estimated in the context of a state-space modelling framework. If one decides to follow this approach, it is important to consider a low-dimensional representation of the spatio-temporal process. It can help us to avoid a over-parameterised model.

It is always relevant the number of EOFs (or PCA) one should consider. One of the simplest ways to select it is just to consider the number of EOFs that account for some desired proportion of overall variance. In particular, due to the number of spatial point is small, $m = 19 \times 3$, it is easier to test different eigenvalues of the EOF in terms of their prediction.

Most EOF analyses in the spatio-temporal context consider spatial EOFs and PC time series. However, one can consider the analogous decomposition in which the EOFs are time-series bases and the projection of the data onto these bases is given by PC spatial fields. One can work with the temporal covariance matrix (averaging over spatial location) or consider the SVD of an $m \times T$ (temporally detrended) data matrix.

Moreover, there are many extensions to the basic EOF analysis presented here such as complex EOFs and multivariate EOFs. These all have specific utilities depending on the type of data and the goal of the analysis. For instance, complex EOFs are used for trying to identify propagating features that account for an important amount of variation in the data. And, multivariate EOFs can considered when multivariate spatial data are observed at the same time points. These methods are described in more detail in Cressie & Wikle (2011) section 5.4.

5.9.1 Estimation in Vector Autoregressive Spatio-Temporal Models via the Method of Moments

In contrast with the traditional vector autoregressive (VAR) time-series applications, the autoregressive process is assumed to correspond directly to the data-generating process. In this context, there are no separate data models and process models. Recall the model defined by the equation

$$\mathbf{Z}_t = \mathbf{M}_t\mathbf{Z}_{t-1} + \boldsymbol{\eta}_t; \qquad \boldsymbol{\eta}_t \sim \ Gau(\mathbf{0}, \mathbf{C}_\eta), \tag{5.44}$$

for $t = 1, \ldots, T$ where it is assumed that $\mathbf{Z}_0$ is known and $\mathbf{Z}_t \equiv (Z_1(\mathbf{s}_1)s, \ldots, Z(\mathbf{s}_m))'$. Estimation of the matrices $\mathbf{M}$ and $\mathbf{C}_\eta$ can be obtained via maximum likelihood, least squares, or the method of moments, see Lütkepohl (2005).

Let us assume $\{\mathbf{Z}_t\}$ has mean zero and is second-order stationary in time. After post-multiplying both sides of 5.44 by $\mathbf{Z}'_{t-1}$ and taking the expectation, one gets,

$$E(\mathbf{Z}_t\mathbf{Z}'_{t-1}) = \mathbf{M}E(\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1}), \tag{5.45}$$

which can be written as

$$\mathbf{C}_Z^{(1)} = \mathbf{M}\mathbf{C}_Z^{(0)}. \tag{5.46}$$

Recall that $\mathbf{C}_Z^{(\tau)}$ is the lag-τ spatial covariance matrix for $\{\mathbf{Z}_t\}$. From 5.46, one has

$$\mathbf{M} = \mathbf{C}_Z^{(1)}(\mathbf{C}_Z^{(0)})^{-1}. \tag{5.47}$$

If one post-multiplies 5.44 by $\mathbf{Z}_t$ take expectations and using 5.47, one has that

$$\mathbf{C}_\eta = \mathbf{C}_Z^{(0)} - \mathbf{M}\mathbf{C}_Z^{(1)'} = \mathbf{C}_Z^{(0)} - \mathbf{C}_Z^{(1)}(\mathbf{C}_Z^{(0)})^{-1}\mathbf{C}_Z^{(1)'}. \tag{5.48}$$

If empirical moments of 5.47 and 5.46 are equated with theoretical moments, one gets

$$\widehat{\mathbf{M}} = \widehat{\mathbf{C}}_Z^{(1)}(\widehat{\mathbf{C}}_Z^{(0)})^{-1}. \tag{5.49}$$

$$\widehat{\mathbf{C}}_\eta = \widehat{\mathbf{C}}_Z^{(0)} - \widehat{\mathbf{C}}_Z^{(1)}(\widehat{\mathbf{C}}_Z^{(0)})^{-1}\widehat{\mathbf{C}}_Z^{(1)'}. \tag{5.50}$$

The equations 5.49 and 5.50 are known as the method-of-moments estimators. In 5.50, the empirical lag-τ covariances matrices, $\mathbf{C}_Z^{(\tau)}$ are calculated as in 4.1. Observe that T needs to be larger than the dimension of $\mathbf{Z}_t$ to ensure that $\widehat{\mathbf{C}}_Z^{(0)}$ is invertible. Reasonable estimates for exploratory data analysis can be obtained from the method-of-moments approach in the context of DSTMs.

Let us project the spatial-mean-centered data onto orthogonal basis functions,

$$\mathbf{\Phi} : \boldsymbol{\alpha}_t = \mathbf{\Phi}'(\mathbf{Z}_t - \widehat{\boldsymbol{\mu}}). \tag{5.51}$$

Now, one can assume that the projected data come from the model $\boldsymbol{\alpha}_t = \mathbf{M}\boldsymbol{\alpha}_{t-\tau} + \boldsymbol{\eta}_t$, and the estimates $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{C}}_\eta$ can be obtained based on the projected data. Then, one can produce forecasts such as $\widehat{\boldsymbol{\alpha}}_{T+\tau} = \widehat{\mathbf{M}}\widehat{\boldsymbol{\alpha}}_T$, with estimated forecast covariance matrix, $\widehat{\mathbf{C}}_\alpha = \widehat{\mathbf{M}}\widehat{\mathbf{C}}_\alpha^{(0)}\widehat{\mathbf{M}}' + \widehat{\mathbf{C}}_\eta$, where $\widehat{\mathbf{C}}_\alpha^{(0)}$ is the empirical estimate of $E(\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t')$. A forecast for $\widehat{\mathbf{Z}}_{T+\tau}$ can be obtained multiplying the forecast $\widehat{\boldsymbol{\alpha}}_{T+\tau}$ by the basis-function matrix and add back the spatial mean:

$$\widehat{\mathbf{Z}}_{T+\tau} = \widehat{\boldsymbol{\mu}} + \mathbf{\Phi}\widehat{\boldsymbol{\alpha}}_{T+\tau}. \tag{5.52}$$

The forecast covariance matrix can be approximated by $\widehat{\mathbf{C}}_Z = \mathbf{\Phi}\widehat{\mathbf{C}}_\alpha\mathbf{\Phi}'$, where the truncation and measurement error when projecting onto the basis functions have been ignored. The parameter estimates in this procedure can give a quick forecast. It can be used as starting values in the state-space EM algorithm that will be described later.

5.9.2 Parameter Estimation via the Expectation–maximisation (EM) Algorithm

One of the most effective ways to estimate unknown parameters in the state-space time-series case is through the EM algorithm.

The basic EM algorithm has been used to find maximum likelihood estimates in the presence of missing data. The *complete data*, W, can be denoted as the union of the observations, W_{obs} and missing data W_{mis}, i.e., $W = (W_{obs} + W_{mis})$. If θ represents unknown parameters, then the complete data density is $[W|\theta]$ and the complete likeli-

hood can be defined as $L(\theta|W) \equiv [W|\theta]$, the loglikelihood is $\log L(\theta|W)$. Let us define $\hat{\theta}^{(i-1)}$ as the parameter estimate at the beginning of the ith iteration of the algorithm; it consists of two steps:

- **E-Step**: Calculate $E(\log L(\theta|W)|W_{obs}, \hat{\theta}^{(i-1)}) \equiv q(\theta|\hat{\theta^{(i-1)}})$.

- **M-Step**: Find θ that maximises $q(\theta|\hat{\theta}^{(i-1)})$, and call this $\hat{\theta}^{(i)}$.

Given a starting value, $\hat{\theta}^{(0)}$, iterate between the E-step and the M-step to obtain a sequence of of estimates, $\{\hat{\theta}^{(i)} : i = 1, 2, \ldots\}$. The algorithm is stopped when a given convergence is reached.

In the state-space version of the EM algorithm, this idea can be extended to calculate unknown parameters. Let us denote by $\mathbf{Z}_{1:T}$ the observations and the unobservable latent process, and $\mathbf{Y}_{0:T}$ the missing data. Denote the parameters by $\mathbf{\Theta} = \{\boldsymbol{\mu}_0, \mathbf{C}_0, \mathbf{C}_\eta, \mathbf{C}_\epsilon, \mathbf{M}\}$, where the observation matrices, $\{\mathbf{H}_t\}$, are often known. The initial distribution is given by $\mathbf{Y}_{0|0} = Gau(\boldsymbol{\mu}_0, \mathbf{C}_0)$ and it is assumed that $\mathbf{C}_\epsilon = \mathbf{C}_{\epsilon,t}$ corresponds to the $m \times m$ error covariance matrix for all possible observation locations, where $m = m_t$ for all t. It can also be assumed no missing observations at each time point, an alternative technique when missing values are observed is proposed in Houseago-Stokes & Challenor (2004). The EM algorithm is based on the complete-data likelihood given by

$$[\mathbf{Z}_{1:T}, \mathbf{Y}_{0:T}|\mathbf{\Theta}] = \left(\prod_{t=1}^{T}[\mathbf{Z}_t|\mathbf{Y}_t]\right)\left(\prod_{t=1}^{T}[\mathbf{Y}_t|\mathbf{Y}_{t-1}]\right)[\mathbf{Y}_0]. \tag{5.53}$$

The EM algorithm 5.10 for a linear DSTM shown below, makes use of the Kalman smoother algorithm to evaluate both the E-step and the M-step.

To evaluate both the E-step and the M-step the use of the Kalman smoother algorithm is required. In addition one needs to obtain the lagged-one smoother variance–covariance matrix,

$$\mathbf{P}_{t,t-1|T} \equiv E((\mathbf{Y}_t - \mathbf{Y}_{t|T})(\mathbf{Y}_{t-1} - \mathbf{Y}_{t-1|T})'|\mathbf{Z}_{1:T}), \quad \text{for } t = T, T-1, \ldots.$$

Convergence can be assessed by considering parameter changes to the log complete-data likelihood, see equation 5.10.

5.10 Implementation using the LAQN data

Some considerations have been taken in order to fit all the DSTM models. One of the main concerns is about the missing values that are present in all the pollutant observations. In particular, to input the missing values in $PM_{2.5}$ the empirical spatial mean is used, see equation 4.21. This is possible since there is at least one observation per site for this pollutant, see figure 2-7. The same procedure was made for the rest of the pollutants. The only situation where this is not possible is for the site HG1, where there are not PM_{10} observations, see figure 2-6. In this case, the empirical temporal mean for the rest of the sites is considered, see equation 4.23. With the aim of generating the EOFs up to 14 March 2016, they can be calculated through a singular value decomposition (SVD). Then the detrended and scaled data matrix, $\tilde{\mathbf{Z}}$, can be obtained.

In this section, a EHM model for a reduced dimension linear DSTM is performed. In this case, the time considered comes from 01 August 2015 to 14 March 2016. The purpose is making predictions for 15 March 2016 based on previous daily $PM_{2.5}$, PM_{10} and NO_2 observations. Let us represent the data as

$$\mathbf{Z}_t \equiv (Z_t(\mathbf{s}_1), \cdots, Z_t(\mathbf{s}_m))',$$

for $m = 19$ monitoring sites, and $t = 1, \cdots, 227 = T$ days. The underlying process can be represented by

$$\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), \cdots, Y_t(\mathbf{s}_n))',$$

where $n = m = 19$. One can assume that the dynamics can be captured by the reduce dimensional process $\{\boldsymbol{\alpha}_t\}$, where,

$$\mathbf{Y}_t = \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\gamma}_t,$$

and $\boldsymbol{\Phi}$ is an $n \times p_\alpha$ matrix of the first p_α EOFs of the empirical spatial covariance matrix of the data. In particular, it is assumed that $\boldsymbol{\gamma}_t \sim iid\ Gau(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$. The purpose is forecasting the smoothed process,

$$\mathbf{Y}_{t+\tau}^P \equiv \Phi\boldsymbol{\alpha}_{t+\tau},$$

given $\mathbf{Z}_{1:t} \equiv \{\mathbf{Z}_1, \cdots, \mathbf{Z}_t\}$, where $\tau = 1$. This lag corresponds to a forecast with 1-day forecast lead time; one should note again that our main interest is forecasting. Since $\{\boldsymbol{\alpha}_t\}$ is assumed to control the dynamics, it is reasonable to take into account only the the smoothed $\{\mathbf{Y}_t^P\}$. If one is interested in spatio-temporal prediction, then the

truncation error, $\boldsymbol{\gamma}_t$, should be considered, and the forecast is $\mathbf{Y}_t \equiv \mathbf{Y}_t^P + \boldsymbol{\gamma}_t$, see Wikle & Cressie (1999); Cressie *et al.* (2010).

Algorithm: Linear DSTM EM Algorithm

Choose initial condition covariance matrix, $\mathbf{C}_0$

Choose starting values: $\hat{\boldsymbol{\Theta}}^{(0)} = \{\hat{\boldsymbol{\mu}}^{(0)}, \hat{\mathbf{C}}_\eta^{(0)}, \hat{\mathbf{C}}_\epsilon^{(0)}, \hat{\mathbf{M}}^{(0)}\}$

repeat $i = 1, 2, \ldots$

1. E-step:
   - Use $\hat{\boldsymbol{\Theta}}^{(i-1)}$ in the Kalman smoother (5.8.2) to obtain $\{\mathbf{Y}_{t|T}^{(i-1)}, \mathbf{P}_{t|T}^{(i-1)}\}$
   - Obtain the lag-one covariance smoother using Kalman smoother output estimates
   - Calculate $\mathbf{P}_{T,T-1|T}^{(i-1)} = (\mathbf{I} - \mathbf{K}_T^{(i-1)}\mathbf{H}_T)\mathbf{M}^{(i-1)}\mathbf{P}_{T-1|T-1}^{(i-1)}$
   - **for** $t = T, T-1, \ldots, 2$ **do**
     $\mathbf{P}_{t-1,t-2|T}^{(i-1)} = \mathbf{P}_{t-1|t-1}^{(i-1)}\mathbf{J}_{t-2}^{(i-1)\prime} + \mathbf{J}_{t-1}^{(i-1)\prime}(\mathbf{P}_{t,t-1|T}^{(i-1)} - \mathbf{M}^{(i-1)}\mathbf{P}_{t-1|t-1}^{(i-1)}\mathbf{J}_{t-2}^{(i-1)\prime}$
   - **end for**
   - Calculate $\mathbf{SS}_{00} = \sum_{t=1} T(\mathbf{P}_{t-1|T}^{(i-1)} + \mathbf{Y}_{t-1|T}^{(i-1)}\mathbf{Y}_{t-1|T}^{(i-1)\prime})$
   - Calculate $\mathbf{SS}_{11} = \sum_{t=1} T(\mathbf{P}_{t|T}^{(i-1)} + \mathbf{Y}_{t|T}^{(i-1)}\mathbf{Y}_{t|T}^{(i-1)\prime})$
   - Calculate $\mathbf{SS}_{10} = \sum_{t=1} T(\mathbf{P}_{t,t-1|T}^{(i-1)} + \mathbf{Y}_{t|T}^{(i-1)}\mathbf{Y}_{t-1|T}^{(i-1)\prime})$
2. M-step:
   - Update: $\hat{\boldsymbol{\mu}}_0^{(i)} = \mathbf{Y}_{0|T}^{(i-1)}$
   - Update: $\hat{\mathbf{M}}^{(i)} = \mathbf{S}_{10}\mathbf{S}_{00}^{-1}$
   - Update: $\hat{\mathbf{C}}_\eta^{(i)} = (1/T)(\mathbf{SS}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{SS}_{10}')$
   - Update:
     $$\hat{\mathbf{C}}_\epsilon^{(i)} = \tfrac{1}{T}\sum_{t=1}^{T}((\mathbf{Z}_t - \mathbf{H}_t\mathbf{Y}_{t|T}^{(i-1)})(\mathbf{Z}_t - \mathbf{H}_t\mathbf{Y}_{t|T}^{(i-1)})' + \mathbf{H}_t\mathbf{P}_{t|T}^{(i-1)}\mathbf{H}_t')$$

until convergence based on differences in:

$$-2\log(L(\boldsymbol{\Theta}^{(i)}|\mathbf{Z}_{1:T}, \mathbf{Y}_{0:T}^{(i)})) = \log(|\hat{\mathbf{C}}_0^{(i)}|) + (\mathbf{Y}_{0|T}^{(i)} - \hat{\boldsymbol{\mu}}_0^{(i)})'\hat{\mathbf{C}}_0^{-1(i)}(\mathbf{Y}_{0|T}^{(i)} - \hat{\boldsymbol{\mu}}_0^{(i)})$$
$$+T\log(|\hat{\mathbf{C}}_\eta^{(i)}|) + \sum_{t=1}^{T}(\mathbf{Y}_{t|T}^{(i)} - \hat{\mathbf{M}}^{(i)}\mathbf{Y}_{t|T}^{(i)})'\hat{\mathbf{C}}_\eta^{-1(i)}(\mathbf{Y}_{t|T}^{(i)} - \hat{\mathbf{M}}^{(i)}\mathbf{Y}_{t|T}^{(i)})$$
$$+T(|\hat{\mathbf{C}}_\epsilon^{(i)}|) + \sum_{t=1}^{T}(\mathbf{Z}_t - \mathbf{H}_t\mathbf{Y}_{t|T}^{(i)})'\hat{\mathbf{C}}_\epsilon^{-1(i)}(\mathbf{Z}_t - \mathbf{H}_t\mathbf{Y}_{t|T}^{(i)}) \ .$$

Now, consider the data model,

$$\mathbf{Z}_t = \mathbf{Y}_t^P + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim iid\ \ Gau(\mathbf{0}, \mathbf{C}_\nu), \tag{5.54}$$

where $\mathbf{Y}_t^P$, which is actually $\boldsymbol{\alpha}_t$, i.e., $\mathbf{Y}_t$ is not featured. In this implementation it is assumed that $\mathbf{C}_\nu = \sigma_\nu^2 \mathbf{I}$. The process model is given by the p_α-dimensional process,

$$\boldsymbol{\alpha}_{t+\tau} = \mathbf{M}\boldsymbol{\alpha}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim Gau(\mathbf{0}, \mathbf{C}_\eta), \qquad t = 1, 2, \cdots \tag{5.55}$$

where the α-process at time $t + \tau$ is conditioned on the α-process at time t to obtain the forecast. If one is interested in following the BHM approach, the parameter model must be specified, see Berliner *et al.* (2000).

Empirical orthogonal functions (EOFs) are a good option for a basis functions to use in this case, since they capture most of the variability. As before, the data from 01 August 2015 to 14 March 2016 will be considered. The EOFs are obtained as described in 5.7, then all unknown parameters are estimated. The first approach is the classical time-series framework based on a vector autoregression and using the method of moments, see section 5.9.1. Then, a state-space framework using the EM algorithm is considered, see section 5.10.

The EOFs can be constructed using only data from 01 August 2015 up to 14 March 2016, they were obtained through a singular value decomposition (SVD). In this example a lead forecast is considered with $p_\alpha = 9$, corresponding to the EOFs in the reduced-dimension linear DSTM given by 5.55. Sensitivity analyses show that keeping $p_\alpha =$ 10,11 and 12 EOFs did not change substantially the forecast results.

After verifying that the time series have mean zero, the covariances can be calculated. Firstly, the lag-0 empirical covariance matrix and the lag-τ empirical cross-covariance matrices are computed, for $\tau = 1$. Then, using the methods of moments described in 5.49 and 5.50, the estimates for $\mathbf{M}$ and $\mathbf{C}_\eta$ can now be calculated from the empirical covariance matrices, see equation 4.24. One-day-ahead forecasts are considered; one projects ahead the EOF coefficients of the time series at the last time point (which corresponds to 14 March 2016) one day into the future. The forecast can be obtained multiplying the forecast $\widehat{\boldsymbol{\alpha}}_{T+\tau}$ by the basis-function matrix and add back the spatial mean.

Spatio-Temporal Validation Statistics

One of the most common scalar validation statistic for continuous-valued spatio-temporal processes is the *mean squared prediction error* (MSPE). It can be defined for the spatio-temporal validation sample $\{Z_v(\mathbf{s}_i; t_j) : j = 1, \ldots, T; i = 1, \ldots, m\}$, and corresponding

predictions $\{\hat{Z}_v(\mathbf{s}_i, t_j)\}$ by

$$MSPE = \frac{1}{Tm}\sum_{j=1}^{T}\sum_{i=1}^{m}\{Z_v(\mathbf{s}_i; t_j) - \hat{Z}_v(\mathbf{s}_i, t_j)\}^2. \tag{5.56}$$

To simplify the notation, it is assumed that the same number of spatial observations for each time period. Of course, different numbers of spatial locations for each time can be accommodated. For now, it is assumed that $\{\hat{Z}(\mathbf{s}_i, t_j)\}$ are predictions based on all of the data, $\mathbf{Z}$. In some circumstances, one could be interested in looking at MSPE for a particular time point, averaged across space, or for a particular spatial location, averaged across time.

The MSPE summary can be thought as an empirical measure of expected squared error loss which, when minimised, results in the S-T kriging predictor. Moreover, the MSPE can be decomposed into a term corresponding to the bias of the predictor plus a term corresponding to the variance of the predictor. It can also be considered the *root mean squared prediction error* (RMSPE), which is the square root of the MSPE, or simply the RMSE (as used below). One advantage is that units of the RMSPE are the same as those of the observations.

One could be interested in a more robust error in term of outliers. It is common to consider the *mean absolute prediction error* (MAPE) defined by

$$MAPE = \frac{1}{Tm}\sum_{j=1}^{T}\sum_{i=1}^{m}|Z_v(\mathbf{s}_i; t_j) - \hat{Z}_v(\mathbf{s}_i, t_j)|. \tag{5.57}$$

5.10.1 Evaluation strategy

In order to assess the accuracy of different methods in forecasting levels of $PM_{2.5}$, we perform a series of studies and evaluate how well they perform at predicting 3-days ahead (with evaluation of 1, 2 and 3 day-ahead forecasts). Due to the nature of the DSTM approach, which does not model the spatial (and between pollutant) dependencies and thus does not lend itself to spatial prediction, in this chapter we focus the evaluation on temporal prediction only.

Three different periods were used to assess the accuracy of forecasts, two in close proximity to other, and the other in a different season. This was to reduce the potential issues that may have arisen if conclusions were drawn based upon a single time period that might have represented a specific set of conditions, e.g. temperature. For each of the three periods, forecasts were produced for three days into the future, using only

data preceding those days, i.e. out-of-sample evaluation. In each case, different lengths of data was used to produce the forecast, ranging from 74 to 227 days, to try and assess the amount of data that might be required to produce accurate forecasts.

The three periods of study are:

1. Forecast: 15, 16, 17 March 2016; Modelling data: 1 August 2015 to 14 March 2016 (227 days); 1 December 2015 to 14 March 2016 (105 days); 1 January 2016 to 14 March 2016 (74 days).

2. Forecast: 26, 27, 28 March 2016; Modelling data: 12 October 2015 to 14 March 2016 (227 days); 12 December 2015 to 14 March 2016 (105 days); 12 January 2016 to 14 March 2016 (74 days).

3. Forecast: 14, 15, 16 October 2015; Modelling data: 1 March 2015 to 13 October 2015 (227 days); 1 July 2015 to 14 March 2016 (105 days); 1 August 2016 to 14 March 2016 (74 days).

In each case, models were fit using the method of moments and an E-M algorithm, as described in (5.9.1) and (5.9.2) respectively and assessed using the RMSPE / RMSE.

5.10.2 Results

DSTM approach assumes data models that have additive Gaussian error and process models that have linear transition structure with additive Gaussian error. In the simplest case, time is discrete and a finite set of spatial locations is assumed. In principle, these types of models are multivariate state-space time series models. Then, sequential prediction and estimation such as filters, smoothers, EM estimations and Bayesian algorithms can be used. A big challenge is the high dimension of the problem. This situation could be due to the amount of data, number of predictions locations or the numbers of parameters to be estimated. Thus, an appropriate parameterisation of the evolution model is required. In case there is no prior knowledge about how to parameterise $\mathbf{M}$, then $\mathbf{M}$ can be estimated in a standard state-space modelling framework.

Here, the values obtained for $\mathbf{C}_0$, $\boldsymbol{\mu}^{(0)}$, $\mathbf{M}^{(0)}$ and $\mathbf{C}_\eta^{(0)}$ from the method of moments are used as the initial values for the EM algorithm, following the suggestion of Balakrishnan *et al.* (2017). In this implementation of the EM algorithm, the tolerance (the smallest change in the log-likelihood, multiplied by 2 across two consecutive iterations of the EM algorithm) was set to 3 and this was reached within 20 iterations. The measurement-error variance was set to $\sigma_\epsilon^2 = 0.1$.

Figure 5-1 shows the estimates of $\{\boldsymbol{\alpha}_t\}$ using both the methods of moments approach and the EM algorithm, the latter using the estimates from the method of moments as initial values. This shows the effects of running the EM algorithm compared to its initial values, i.e. the addition of iteratively using the data (through the log-likelihood) to estimate the parameter values.

The estimates of the evolution matrix for the method of moments and EM algorithm can be seen in figure 5-2 and the observed patterns within the evolution matrices look similar. One the advantages about the DSTM approach is that forecasting using this EOF-reduced model is straightforward. One can take the coefficients at the final time point, $\{\boldsymbol{\alpha}_t\}$, propagate those forward, and re-project onto the original space. For example, $\hat{\boldsymbol{\mu}}+\mathbf{M}^2\boldsymbol{\alpha}_t$ gives a two-day forecast, these results are not included here. Matrix powers represent multiple matrix multiplications in this case.

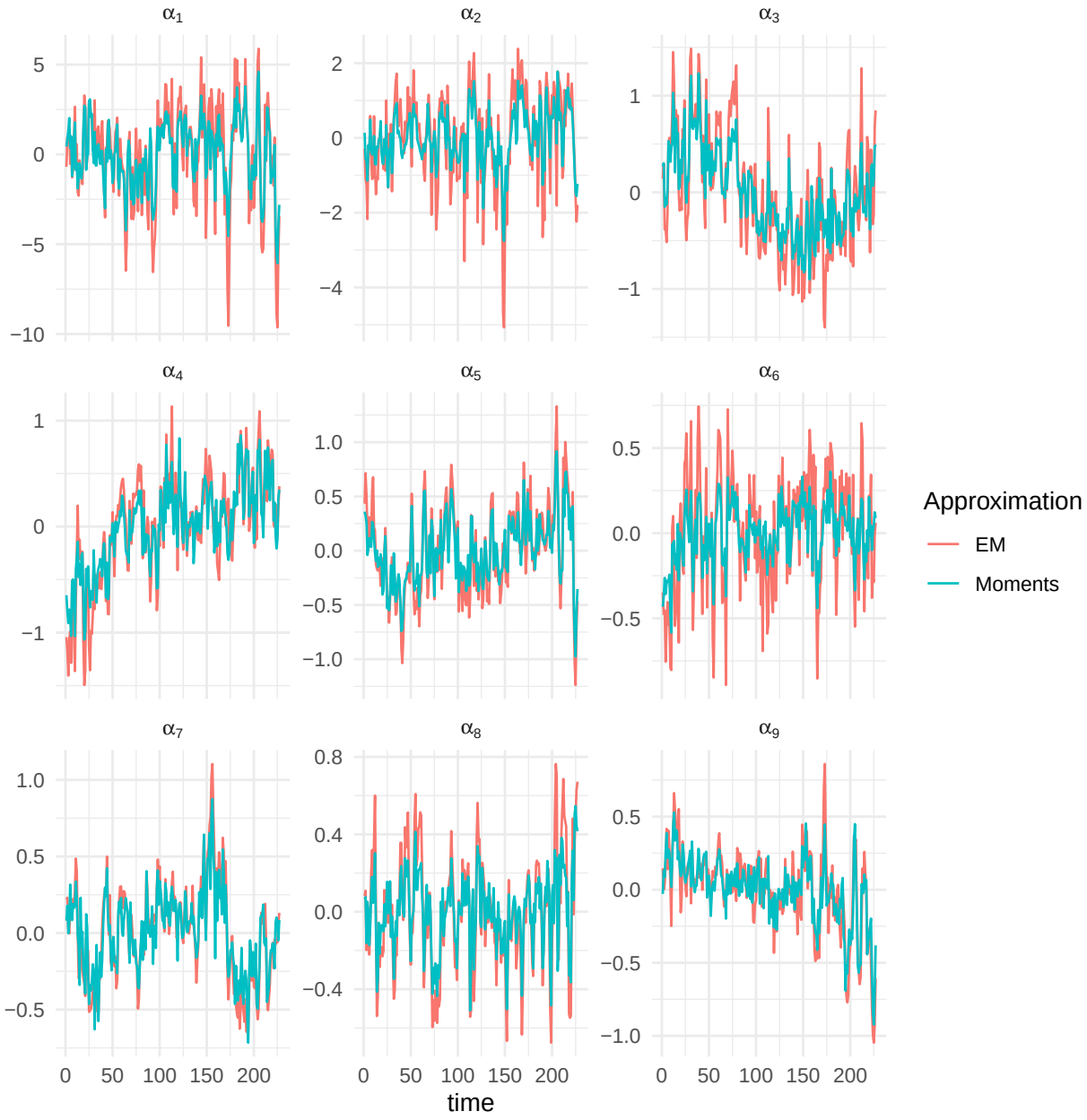

Figure 5-1: Reduced dimension linear DSTM results for $PM_{2.5}$ concentrations from 01 August 2015 to 14 March 2016. Estimates of $\alpha_{i,t}, \quad i = 1, \ldots, 9$, using the method of moments (light blue line) and the EM algorithm (light red line).

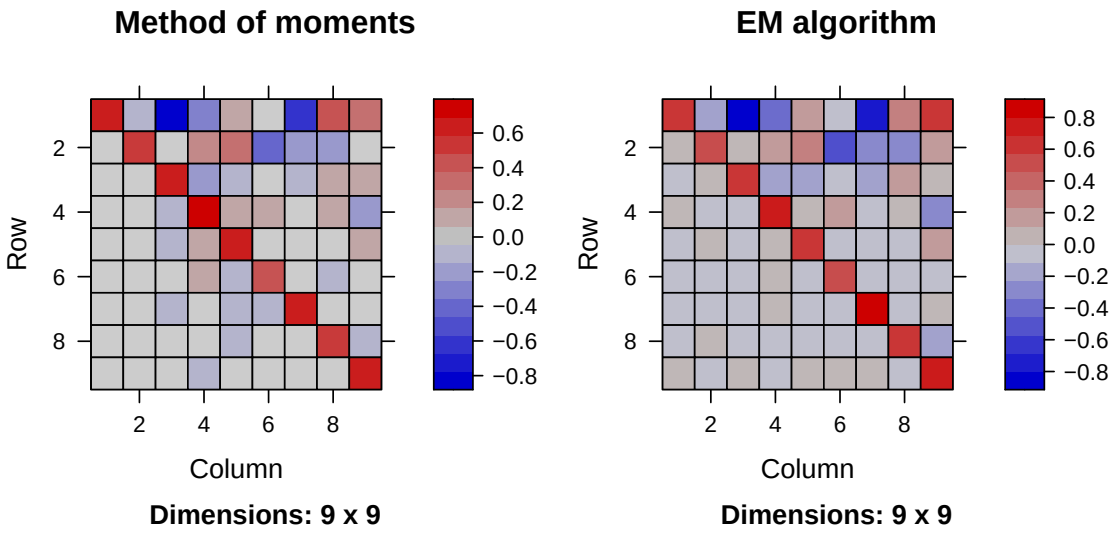

Figure 5-2: Reduced dimension linear DSTM results for $PM_{2.5}$ concentrations from 01 August 2015 to 14 March 2016. The estimate of the one-steps-ahead evolution operator using the method of moments (left) and the estimate of a one-step-ahead evolution operator using the EM algorithm (right).

A study is proposed in order to evaluate the accuracy of the DSTM models in terms of prediction for three different scenarios. The forecast for 15, 16, and 17 March will be produced under the DSTM framework using different days of data as input. The analysis consists in analysing the results for three different periods of time. The first interval corresponds to the period from 01 August 2015 to 14 March 2016 (227 days). This period will be used to fit a method of moments model as it was explained in section 5.9.1 and an EM-algorithm model such as it is described in the algorithm 5.9.2. Due the simplicity of the method of moments estimators, they are easy to implement and it is an alternative for a preliminary approximation. The EM algorithm carries out maximum likelihood estimation in a state-space model is implemented.

In a similar way, these two approaches will be considered for the period from 01 December 2015 to 14 March 2015 (105 days). Finally, the third period goes from 01 January 2016 to 14 March 2016 (74 days). In addition, due to the flexibility of the DSTM framework, it can be extended to analyse not just $PM_{2.5}$ observations. As it was explained in section 5.5.1, the model can be adapted to consider PM_{10} and NO_2 data simultaneously. It is a multivariate approach in which $PM_{2.5}$ predictions take into account the correlations with other pollutants. One of the purposes of this study is to see if there is an improvement of the forecast when more data is added. The comparison is made with actual values of $PM_{2.5}$ that were previously removed when fitting the model. The *root of the mean squared error*, described in equation 5.56 is used to measure the error.

The forecast for 15, 16 and 17 March 2016 using the method of moments and EM-algorithm can be seen in the table 5.1. These results are associated to the model that considers all the pollutants, $PM_{2.5}$, PM_{10} and NO_2. From now, some conventions will be followed for the column names in all the tables. The site column denotes the code from LAQN (see table 2.1) and the real column represents the actual value at each site. Every column with the name MM will represent the results for the method of moments, meanwhile EM will denote the results for the EM-algorithm. The multicolumn 7 months refers to the period from 01 August 2015 to 14 March 2016. On the other hand, the period from 01 December 2015 to 14 March 2015 is represented by the 3 months multicolumn. And, the 2 months multicolumn corresponds to the period from 01 January 2016 to 14 March 2016. Finally, the last row for each table shows the RMSE for each predicted time. The results for the model that just considers $PM_{2.5}$ data as input are shown in the table 5.2. The same conventions about column names will follow for the rest of the tables.

There are some general observations for the tables 5.1 and 5.2. In terms of RMSE, there

is not a large difference when using 7, 3 and 2 months of data. In the same way, a small difference is observed between MM and EM for both tables. The best prediction for 15 March 2016 is the EM-algorithm for 3 months (just $PM_{2.5}$ data), the error is similar to the MM method for the same period of time. The best prediction for 16 March 2016 is observed when fitting the EM-algorithm for 3 months (all pollutants). On the other hand, the best result for the 17 March 2016 is the EM-algorithm for 7 months (just $PM_{2.5}$ data). There is a significant variation in accuracy for the prediction for different sites. For instance, the biggest difference in error for 15 March 2016 is found in the site HG1. As one would expect in this type of model, the RMSE rises when more days ahead are predicted.

In order to compare the previous results with another time interval, two analogous studies are proposed. In the first case the result forecast is done for 26, 27 and 28 March 2016. The three periods of time starts on 12 October 2015 (227 days), 12 December 2015 (105 days) and 12 January 2015 (74 days). In particular, all the models are fitted using up to 25 March 2016 data. The results considering the three pollutants and just $PM_{2.5}$ data are found in the tables 5.3 and 5.4 respectively.

The difference between using 7, 3 and 2 months of data is slightly larger than the first time interval in terms of RMSE. In general, a small difference is observed between MM and EM for both tables. The best prediction for 26 March 2016 is the INLA model for 2 months, the error is similar to the DSTM model for 7 months. The best prediction for 27 March 2016 is observed when fitting the INLA model for 2 months. Finally, the best result for the 28 March 2016 is the INLA model for 2 months. There is also an important variation in accuracy for the prediction for different sites. In general, RMSEs are larger for this study than those ones from the first analyses.

In the second case the result forecast is done for 14, 15 and 16 October 2015. The three periods of time starts on 01 March 2015 (227 days), 01 July 2015 (105 days) and 01 August 2015 (74 days). In particular, all the models are fitted using up to 13 October 2015 data. The results considering the three pollutants and just $PM_{2.5}$ data are found in the tables 5.5 and 5.6 respectively. In terms of RMSE, the difference between using 7, 3 and 2 months of data is slightly larger than the previous results. In general, a small difference is observed between MM and EM for both tables. The best prediction for 14 October 2015 is the MM for 7 months (just $PM_{2.5}$ data), the error is similar to the EM method for the same period of time. The best prediction for 15 October 2015 is observed when fitting the MM for 7 months (just $PM_{2.5}$ data). Finally, the best result for the 16 October 2015 is the MM for 2 months (just $PM_{2.5}$ data). In contrast to the previous tables, the RMSEs are larger for one day ahead than for two

days ahead forecast. There is also an important variation in accuracy for the prediction for different sites. In general, RMSEs are larger for this study than those ones from the previous analyses.

5.11 Discussion

In this chapter, the ability of the DSTM to produce accurate forecasts of $PM_{2.5}$ is assessed. Forecasts are produced for three different time periods considering the method of moments and an EM algorithm. The focus here is on the ability to predict accurately over time, rather than space, due to the nature of the DSTM approach (see Section 5.10.1).

The forecast for the method of moments and the EM algorithm are, in general, similar. This is not surprising, since the estimates $\{\boldsymbol{\alpha}_t\}$ and the estimates of the evolution matrices are similar. The DSTM approach allows one to analyse $PM_{2.5}$ data and more pollutants (PM_{10}, NO_2) simultaneously, but based on the findings of the evaluation studies, there is no general pattern of improvement when including additional pollutants (in the ability to forecast $PM_{2.5}$.) Also, there is not much improvement when additional data (over time) is used. It should be noted that even the minimum length of time is 74 days, which is quite substantial.

Although the DSTM approach allows a computationally efficient approach to modelling large spatio-temporal datasets, including multiple pollutants, one of the major disadvantages of DSTM is that this approach does not allow generation of spatial prediction. This is an important limitation if one is interested in producing spatial prediction and maps of pollution. DSTMs capture correlations non-spatially via the transition matrix, while other methods, such as those considered in the following chapters (Spatio-temporal Kriging in Chapter 6 and a Bayesian Spatio-temporal model in Chapter 7), incorporate it within covariance matrices. As we move onto those Chapters, we will explore the advantages of such methods in terms of spatial prediction and develop a series of evaluation studies to assess this.

Another disadvantage of DSTM EM algorithm models is that they do not take the uncertainty of the parameter estimates directly into account. In principle, the BHM formulation can include Gaussian structures trough MCMC. Inference for DSTM is evolving quickly; for a MCMC-related implementation see Cressie & Wikle (2011).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	19.1875	16.9955	19.1226	18.5460	19.9941	15.8328	17.2844
	BX1	15.7917	15.7551	17.1619	16.2077	16.8875	14.0037	14.3084
	CD1	21.3875	19.5159	21.1778	21.3691	22.4825	19.1377	20.3011
	CD9	26.4237	23.6159	25.2070	23.3119	24.1407	20.7242	21.7778
	CT3	22.7000	20.4973	21.7511	20.2717	20.9220	19.5258	20.5908
	GN0	19.8932	18.7391	20.4904	21.6921	22.9310	20.4003	22.1814
	GN2	20.6132	20.4472	22.6599	20.8189	22.0068	18.1684	19.7476
	GN3	20.8812	22.1929	23.8452	22.7302	23.7345	20.1329	21.0416
	GR4	16.2870	16.0993	17.6597	16.6452	17.5135	14.7066	15.7098
15/03/16	GR9	20.7735	21.8481	24.4626	22.0426	23.5738	19.3270	20.8842
	HG1	9.2000	11.9135	12.0039	9.6551	9.4238	9.0489	9.1123
	HR1	11.4000	9.9603	10.1866	12.3415	12.5880	11.4907	12.0112
	HV1	15.2052	14.5228	15.2813	17.2218	17.8268	15.3011	16.1128
	KC1	19.0875	16.8333	18.5658	18.2281	19.3327	15.5754	16.7305
	KC2	22.6098	21.3092	22.8348	21.4582	22.2415	18.6901	19.3909
	LH0	18.2871	16.5415	18.4848	17.2814	18.3877	15.7367	17.3623
	LW2	22.4770	23.0902	24.2884	24.4631	25.0059	21.8462	22.4256
	ST5	20.2300	20.6047	22.1305	19.6192	20.0682	17.2551	17.7591
	TH4	11.4372	11.4220	11.4382	11.4372	11.4372	11.4372	11.4372
	RMSE		1.5484	1.6178	1.3984	1.7021	2.4810	1.8278
	BL0	15.7714	13.4194	16.2578	13.5611	16.8833	12.6030	14.7321
	BX1	16.1250	13.2610	14.8491	12.3188	14.1600	10.5883	11.5437
	CD1	16.8542	17.3586	19.3303	16.8710	19.4903	16.0620	17.4487
	CD9	23.6964	21.6680	23.3777	19.5872	21.9752	18.7440	20.1530
	CT3	18.7500	18.0262	19.9705	16.6768	18.4097	16.7780	17.9580
	GN0	18.7205	16.3717	18.4500	17.1726	20.6256	17.0689	19.3751
	GN2	19.2868	17.0284	19.5807	15.8785	18.8793	14.9690	16.8581
	GN3	18.8502	19.7371	21.7622	18.2856	20.8156	17.2042	18.2657
	GR4	15.7458	14.3105	15.8164	12.7855	14.6297	11.7449	12.8454
16/03/2016	GR9	19.4564	17.9213	21.0512	16.5350	20.1514	15.0047	17.6743
	HG1	9.2000	13.5560	12.6014	10.4484	9.5341	10.2170	9.5626
	HR1	16.0500	9.9901	10.0219	12.0527	12.6845	11.8574	12.2652
	HV1	15.3946	13.1937	14.2007	13.9638	15.7917	13.4759	14.3783
	KC1	15.4792	14.4699	16.5458	13.8883	16.4921	12.9051	14.2718
	KC2	19.2250	18.9255	20.9065	17.7539	19.8105	16.5438	17.3476
	LH0	15.3568	13.1718	15.9752	13.2031	15.7311	12.7280	14.6694
	LW2	19.4649	21.1965	22.8953	20.9257	22.7108	19.9266	20.2717
	ST5	17.9700	18.9106	20.5227	16.5117	17.9081	15.1870	15.5989
	TH4	11.4372	11.8245	11.5274	11.4372	11.4372	11.4372	11.4372
	RMSE		2.3497	2.2094	2.3549	1.5920	3.0832	2.0658
	BL0	32.8500	9.2357	13.7242	9.5661	14.5320	10.8607	13.3662
	BX1	26.4250	9.6014	12.9596	8.9521	12.5878	8.6092	10.1608
	CD1	30.9917	13.2813	17.3864	12.2421	17.4717	13.4888	15.9740
	CD9	34.9552	18.6285	21.7243	16.5182	20.4129	17.8380	19.4215
	CT3	28.3100	15.2818	18.3403	13.1921	16.8907	14.1148	16.4149
	GN0	28.3283	12.1558	16.2962	12.7503	18.7281	13.6877	17.2085
	GN2	31.9742	12.3051	16.8896	11.1999	16.7026	12.3314	15.1781
	GN3	30.1591	15.3973	19.6688	14.0363	18.9241	14.7763	16.8756
	GR4	25.6125	10.6610	14.0569	8.8015	12.8912	9.2374	11.3533
17/03/2016	GR9	30.7078	13.1676	18.0737	12.0989	17.6093	12.6659	15.7367
	HG1	9.2000	12.9041	12.8769	9.9806	9.7528	10.3096	9.9066
	HR1	28.4083	8.6638	9.7051	10.6969	12.4304	12.0138	12.4321
	HV1	24.9809	10.1367	13.0611	10.7410	14.4051	11.7348	13.3983
	KC1	27.7333	10.4892	14.5125	9.6857	14.5353	10.5270	12.8699
	KC2	19.2250	15.8411	19.1541	14.4015	18.2194	15.2272	16.5209
	LH0	29.1291	9.2442	13.5955	9.2030	13.7113	10.0834	12.8794
	LW2	30.7497	17.0815	21.2422	16.9785	21.1049	17.4142	19.1666
	ST5	27.4300	15.3161	18.8469	12.9946	16.4715	12.8188	14.4221
	TH4	11.4372	11.9251	11.6362	11.4372	11.4372	11.4372	11.4372
	RMSE		15.6442	13.1362	16.2087	12.8087	15.4644	13.4706

Table 5.1: Evaluation of applying the DSTM model to data from 01/08/15 to 14/03/16 considering multi-pollutant data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 15, 16 and 17 of March, 2016. All comparisons were made in the original scale (μg/m^3).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	19.1875	16.9055	18.1135	17.0367	17.0459	17.7269	17.7342
	BX1	15.7917	17.8562	18.5204	17.2246	16.4736	17.4044	16.0115
	CD1	21.3875	19.9104	20.8240	20.9667	20.7737	21.6964	20.8164
	CD9	26.4237	24.5893	26.5421	24.5531	24.6731	24.4015	24.1680
	CT3	22.7000	20.3207	21.1271	20.8942	20.7284	21.0109	20.5151
	GN0	19.8932	18.3581	18.8980	20.6001	20.0295	20.2397	18.3337
	GN2	20.6132	20.8752	22.1253	20.3456	19.8165	20.3725	18.8947
	GN3	20.8812	21.4934	22.0121	22.0507	21.7488	22.5120	21.8269
	GR4	16.2870	15.7569	16.5941	16.5130	16.1965	16.1758	15.5926
15/03/16	GR9	20.7735	21.5926	23.7338	21.4846	21.6694	21.3076	20.7234
	HG1	9.2000	10.5051	9.9723	10.1493	9.6506	9.4235	9.2067
	HR1	11.4000	10.6966	10.3661	10.7542	10.2392	11.3368	10.2043
	HV1	15.2052	16.3250	16.2776	17.7039	17.1563	18.3028	17.4372
	KC1	19.0875	16.7721	17.7376	17.8065	17.7018	18.0294	17.7151
	KC2	22.6098	21.6725	23.3273	21.5744	21.3949	21.8005	21.4080
	LH0	18.2871	16.3493	17.3058	16.1661	15.8511	16.0701	15.4319
	LW2	22.4770	24.0987	24.3674	24.0152	23.8502	25.3068	25.0499
	ST5	20.2300	20.4123	21.5647	20.0966	19.8099	20.2485	19.7566
	TH4	11.4372	11.4139	11.4981	11.4372	11.4372	11.4372	11.4372
	RMSE		1.4607	1.3821	1.3233	1.2844	1.4374	1.5191
	BL0	15.7714	12.6705	15.8697	12.7501	14.3915	13.6943	14.9274
	BX1	16.1250	13.8634	16.3606	12.3787	13.8724	13.0768	13.4314
	CD1	16.8542	16.2824	18.8083	16.9359	18.2312	17.9038	18.2032
	CD9	23.6964	20.8786	24.4993	20.9362	22.4990	20.7859	21.6388
	CT3	18.7500	17.6594	19.4297	18.1748	18.7567	17.6546	18.2137
	GN0	18.7205	15.0730	17.1031	16.5215	17.6748	16.9742	16.1966
	GN2	19.2868	17.0524	19.6208	16.0984	17.0420	16.1592	16.0515
	GN3	18.8502	18.0737	20.2552	18.2685	19.5227	18.9667	19.3985
	GR4	15.7458	12.9243	14.8898	12.9858	13.9975	12.9688	13.3977
16/03/16	GR9	19.4564	16.9109	20.8761	16.3829	18.3477	16.6397	17.7039
	HG1	9.2000	10.7819	10.2556	10.6956	9.8397	9.7831	9.2439
	HR1	16.0500	10.3879	10.1867	10.8266	9.9587	11.2521	9.9420
	HV1	15.3946	13.1222	14.8340	14.2667	15.1640	14.8662	15.1005
	KC1	15.4792	13.2441	15.7524	13.6797	15.2105	14.3414	15.1383
	KC2	19.2250	18.4954	21.4150	18.2531	19.4702	18.5785	19.1699
	LH0	15.3568	12.3788	14.9528	12.5709	13.5303	12.7381	13.1636
	LW2	19.4649	20.9640	22.6030	20.8721	21.9504	21.9437	22.6806
	ST5	17.9700	17.3422	19.6985	16.6979	17.8142	17.0208	17.5722
	TH4	11.4372	11.4644	11.5852	11.4372	11.4372	11.4372	11.4372
	RMSE		2.4526	1.8657	2.4046	1.8976	2.2020	2.2107
	BL0	32.8500	9.6771	14.3515	9.2848	12.8429	10.3333	13.0602
	BX1	26.4250	10.5398	14.7691	9.1095	12.5313	9.5445	11.7431
	CD1	30.9917	12.9976	17.4548	12.8165	16.7054	13.8565	16.4578
	CD9	34.9552	17.7847	23.2599	17.4574	20.8401	17.1877	19.8929
	CT3	28.3100	14.4450	18.2179	14.0931	17.5019	14.2035	16.6532
	GN0	28.3283	12.0690	15.8630	12.4917	15.8980	13.2043	14.7868
	GN2	31.9742	12.8113	18.0087	11.5534	15.2287	11.9953	14.2042
	GN3	30.1591	14.9010	19.0076	14.3830	18.2110	15.1843	17.7555
	GR4	25.6125	10.1531	13.7937	9.3811	12.7133	9.6380	11.9596
17/03/16	GR9	30.7078	13.2650	19.0660	12.1456	16.1925	12.3818	15.6785
	HG1	9.2000	11.0325	10.5195	10.6787	9.9529	10.0625	9.3077
	HR1	28.4083	9.7344	10.0842	10.1211	9.7282	10.9524	9.7748
	HV1	24.9809	10.3387	13.6946	10.8636	14.0085	11.6559	13.5466
	KC1	27.7333	10.2814	14.4549	9.9087	13.7251	10.6591	13.4434
	KC2	19.2250	15.4817	20.2362	15.0222	18.2468	15.2898	17.6334
	LH0	29.1291	9.2047	13.3570	9.0014	12.1183	9.5690	11.6685
	LW2	30.7497	17.3533	21.3227	17.1125	21.0899	18.2313	21.0439
	ST5	27.4300	14.2953	18.4927	13.2057	16.7591	13.5563	16.0919
	TH4	11.4372	11.6486	11.6846	11.4372	11.4372	11.4372	11.4372
	RMSE		15.6304	11.9422	16.0316	13.0422	15.5008	13.4923

Table 5.2: Evaluation of applying the DSTM model to data from 01/08/15 to 14/03/16 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 15, 16 and 17 of March, 2016. All comparisons were made in the original scale ($\mu g/m^3$).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	6.6083	7.0618	7.7453	7.6697	7.957	6.9437	7.1474
	BX1	4.6583	6.1615	6.5411	6.7157	7.0661	6.2859	6.3095
	CD1	8.2833	10.4126	10.7708	10.6099	10.8914	10.2949	10.3841
	CD9	9.3778	13.2639	13.2626	12.882	13.1902	12.5415	12.7355
	CT3	7.05	11.3483	11.6492	11.1014	11.2813	10.7331	10.5527
	GN0	11.5432	8.8028	9.0811	8.8398	9.2751	9.2477	9.4042
	GN2	6.4976	7.6507	7.9111	8.1127	8.4788	7.7592	7.96
	GN3	10.6922	11.5106	11.7129	11.8743	12.1292	11.6878	11.5921
	GR4	6.5909	6.8132	6.9002	7.3875	7.5722	7.1382	7.0006
26/03/2016	GR9	8.0627	8.4255	8.8126	9.0618	9.4174	8.5814	9.0523
	HG1	9.2	9.62	8.6652	9.474	9.4376	9.3463	8.6602
	HR1	6.6458	8.5307	8.787	9.3552	9.7032	8.9098	9.476
	HV1	7.724	8.049	8.3388	8.4257	8.6822	8.0025	7.8701
	KC1	6.025	7.4052	7.6981	7.8099	8.0319	7.7365	7.7114
	KC2	19.225	12.1151	12.4201	12.0733	12.3186	11.6065	11.6831
	LH0	6.6806	6.1576	6.6728	6.9922	7.2406	6.6182	6.7055
	LW2	14.2416	16.0409	16.5843	16.0435	16.1868	15.6538	15.7059
	ST5	9.36	10.9993	11.006	11.2921	11.4683	11.0965	11.0502
	TH4	11.4372	11.6306	11.3746	11.4372	11.4372	11.4372	11.4372
	RMSE		2.4365	2.5088	2.5312	2.6296	2.4063	2.4354
	BL0	3.5455	7.5647	8.2606	8.6978	8.5342	7.7783	7.8239
	BX1	1.5696	6.6605	6.9621	7.527	7.548	6.6576	6.8654
	CD1	4.6833	11.0747	11.4084	11.7736	11.5052	11.2806	11.1163
	CD9	6.0587	14.3332	13.9674	14.32	14.1101	14.0501	13.8789
	CT3	5.23	12.3096	12.2537	12.241	11.8944	11.6471	11.3745
	GN0	7.0242	9.2862	9.7384	9.8666	10.0116	10.2298	10.1689
	GN2	3.5175	8.309	8.5921	9.2692	9.2098	8.8171	8.804
	GN3	6.5583	12.0651	12.3236	12.9877	12.7599	12.6897	12.4908
	GR4	2.8957	7.3268	7.3611	8.3209	8.0496	7.8091	7.5901
27/03/2016	GR9	4.3177	9.1424	9.3822	10.1342	10.0411	9.3302	9.6047
	HG1	9.2	9.5805	8.9968	9.7953	9.6678	9.7165	9.1531
	HR1	4.0167	8.33	9.0707	9.6197	10.0129	9.2358	9.7211
	HV1	3.305	8.5193	8.8397	9.5151	9.3647	8.8753	8.7624
	KC1	3.3167	7.9453	8.2675	8.9236	8.6483	8.7308	8.4516
	KC2	19.225	12.916	13.0178	13.1582	12.9485	12.7631	12.5774
	LH0	3.6658	6.7436	7.2435	7.9893	7.8261	7.2913	7.2593
	LW2	7.4832	16.5354	17.2509	17.3651	16.8961	16.8387	16.5718
	ST5	6.11	11.7668	11.5615	12.4405	12.1144	11.9091	11.7494
	TH4	11.4372	11.62	11.3288	11.4372	11.4372	11.4372	11.4372
	RMSE		5.2986	5.5134	5.9403	5.7956	5.5322	5.4766
	BL0	3.2136	8.102	8.5343	9.4888	8.8238	8.9582	8.3819
	BX1	2.2177	7.2529	7.324	8.3583	7.855	7.6679	7.3034
	CD1	4.487	11.7214	11.7609	12.6345	11.8255	12.4008	11.4772
	CD9	6.6365	15.1018	14.5116	15.4819	14.6855	15.8895	14.5898
	CT3	3.71	13.0851	12.6452	13.1796	12.2646	12.8113	11.7152
	GN0	5.5313	9.8807	10.1387	10.9437	10.4084	11.4367	10.6144
	GN2	2.7521	9.1358	9.0755	10.3082	9.6032	10.2006	9.328
	GN3	5.9556	12.8383	12.7516	14.0594	13.12	13.9987	12.8813
	GR4	3.0714	7.879	7.7225	9.1363	8.3382	8.6851	7.8274
28/03/2016	GR9	3.8921	9.8183	9.8173	11.1032	10.4041	10.5082	10.2588
	HG1	9.2	9.5906	9.3026	9.8115	9.7685	9.9406	8.9528
	HR1	3.2217	8.3659	9.1938	9.8043	10.0339	9.6864	10.0025
	HV1	3.9784	9.1538	9.1806	10.3992	9.7045	10.1415	9.1623
	KC1	3.0136	8.4846	8.6221	9.7934	8.9791	9.8386	8.7784
	KC2	19.225	13.5715	13.4012	14.1347	13.3702	14.1963	13.1124
	LH0	2.1352	7.1889	7.6025	8.7897	8.115	8.2069	7.6353
	LW2	9.7832	17.3759	17.4795	18.4033	17.1471	17.9791	16.86
	ST5	6.8557	12.4931	11.9677	13.4318	12.4111	12.9196	12.0353
	TH4	11.4372	11.6732	11.3318	11.4372	11.4372	11.4372	11.4372
	RMSE		5.8792	5.8789	6.7702	6.1452	6.5579	5.8922

Table 5.3: Evaluation of applying the DSTM model to data from 12/08/15 to 25/03/16 considering multi-pollutant data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 26, 27 and 28 of March, 2016. All comparisons were made in the original scale (μg/m^3).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	6.6083	9.3269	9.7586	8.4385	8.9586	12.3212	11.9904
	BX1	4.6583	7.3424	7.4708	6.5420	6.6143	11.0902	9.9655
	CD1	8.2833	11.2528	11.8133	11.1006	11.3853	15.9447	15.1462
	CD9	9.3778	14.5505	14.8912	15.4323	15.4974	17.9779	17.5204
	CT3	7.0500	11.5345	11.8759	11.2454	10.9867	14.7218	14.3958
	GN0	11.5432	12.2966	11.6907	10.7871	11.0121	15.7167	14.9639
	GN2	6.4976	11.8605	10.6318	9.0427	9.1881	14.2850	13.6899
	GN3	10.6922	13.4746	13.0753	12.5011	12.5168	16.6105	15.8849
	GR4	6.5909	7.7391	7.5979	7.5249	7.4382	10.7427	10.0642
26/03/16	GR9	8.0627	10.3312	10.7488	9.4499	9.9334	13.8188	13.1089
	HG1	9.2000	10.9519	9.0141	9.9377	9.1258	9.1024	9.3169
	HR1	6.6458	9.6824	10.2298	9.1100	9.6988	11.7348	12.4575
	HV1	7.7240	9.2075	9.0963	9.0559	9.0491	12.7602	12.2031
	KC1	6.0250	9.1381	9.1958	8.5846	8.7196	11.3606	10.8952
	KC2	19.2250	15.7004	14.8438	14.0402	13.9028	18.4565	17.5405
	LH0	6.6806	8.6046	8.1460	7.3999	7.5674	10.8555	10.5295
	LW2	14.2416	18.0182	17.9306	15.7636	15.8879	20.5515	20.1472
	ST5	9.3600	11.5497	11.3848	11.1540	11.0857	14.7664	14.1130
	TH4	11.4372	10.6804	10.9676	11.4372	11.4372	11.4372	11.4372
	RMSE		3.0240	3.0702	2.6166	2.7157	5.5997	5.1560
	BL0	3.5455	11.3546	10.7237	9.0522	9.4204	19.6554	10.2646
	BX1	1.5696	9.5626	8.4076	7.4777	7.1021	21.3670	8.2097
	CD1	4.6833	13.8735	13.0006	12.2049	12.1062	22.7638	12.1948
	CD9	6.0587	16.6067	15.7025	15.8154	15.9691	23.7347	17.1594
	CT3	5.2300	13.3637	12.6760	12.4732	11.8144	21.1007	12.8228
	GN0	7.0242	14.3197	12.7971	11.4118	11.4223	22.0836	11.3728
	GN2	3.5175	14.1325	11.7878	9.9805	9.9039	22.9209	10.6896
	GN3	6.5583	15.8945	14.2782	13.7530	13.2567	23.5402	13.6736
	GR4	2.8957	9.9519	8.6032	8.5566	8.0311	16.9181	8.6298
27/03/16	GR9	4.3177	12.6906	11.7395	10.2629	10.4772	21.8532	11.2697
	HG1	9.2000	11.9040	9.6720	9.8506	9.2900	9.4367	10.3582
	HR1	4.0167	12.1284	11.2890	9.7394	10.3674	17.8878	10.6520
	HV1	3.3050	11.5083	10.1771	10.1291	9.7469	21.5300	10.5678
	KC1	3.3167	11.3539	10.2824	9.5700	9.2928	18.2049	9.7883
	KC2	19.2250	16.4307	15.3737	14.2196	14.1987	19.8606	14.8079
	LH0	3.6658	10.7819	9.3377	8.2496	8.1347	18.3076	8.8350
	LW2	7.4832	20.1714	19.1660	17.5412	16.8964	27.2587	16.7251
	ST5	6.1100	13.7388	12.4045	12.5590	11.8465	21.7012	12.1583
	TH4	11.4372	10.8600	10.9257	11.4372	11.4372	11.4372	11.4372
	RMSE		8.0873	7.0230	6.3133	6.1362	15.4472	6.6370
	BL0	3.2136	11.9295	11.2846	9.1633	9.4120	21.4420	14.7872
	BX1	2.2177	10.5346	8.8895	7.8015	7.3937	25.0180	13.3184
	CD1	4.4870	14.9152	13.6662	12.6476	12.2939	24.9935	17.8394
	CD9	6.6365	17.0616	16.3289	15.6955	16.0084	24.8157	20.6405
	CT3	3.7100	13.8518	13.1351	12.8024	12.2079	22.5194	16.4550
	GN0	5.5313	14.9256	13.2219	11.8027	11.5160	23.3720	17.3405
	GN2	2.7521	14.3254	12.2681	10.3942	10.1373	24.7011	16.9114
	GN3	5.9556	16.6473	14.8091	14.1114	13.4995	24.7803	18.3672
	GR4	3.0714	10.9724	9.0967	8.9757	8.3128	19.1402	12.3191
28/03/16	GR9	3.8921	13.8461	12.3496	10.6825	10.6964	24.4401	16.2717
	HG1	9.2000	12.0928	9.7738	9.8078	9.3878	9.4332	9.4941
	HR1	3.2217	12.9732	11.8403	10.3924	10.4797	19.2129	14.4450
	HV1	3.9784	11.9452	10.6967	10.3202	9.9374	21.8637	15.0221
	KC1	3.0136	12.2279	10.8214	9.8976	9.4511	20.4769	13.4334
	KC2	19.2250	15.9541	15.6545	14.0129	14.2070	19.4237	18.9582
	LH0	2.1352	11.4653	9.8311	8.6149	8.3306	20.7512	13.0573
	LW2	9.7832	20.5795	19.6483	17.9966	17.1805	27.8674	22.0793
	ST5	6.8557	14.6188	12.9194	13.0030	12.1766	23.0367	16.3364
	TH4	11.4372	11.0498	10.9790	11.4372	11.4372	11.4372	11.4372
	RMSE		8.8556	7.6074	6.7158	6.4216	17.1799	10.8671

Table 5.4: Evaluation of applying the DSTM model to data from 12/08/15 to 25/03/16 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 26, 27 and 28 of March, 2016. All comparisons were made in the original scale ($\mu g/m^3$).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	6.3750	9.8093	9.8467	9.4774	9.4970	10.5813	10.7827
	BX1	5.2333	10.0070	10.1131	10.1047	9.9440	11.0081	11.1283
	CD1	10.8458	14.5879	14.6992	15.3375	15.3770	15.3136	15.6446
	CD9	10.6696	17.4710	17.5440	21.0252	21.1104	22.3027	22.4075
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	11.0350	13.0261	12.9894	13.1889	13.2280	13.9707	14.2529
	GN2	6.8770	13.1919	13.3747	13.0604	12.7540	14.5223	14.2850
	GN3	10.2575	15.5134	15.6451	16.6747	16.6735	17.1379	17.3924
	GR4	5.7458	10.3325	10.7142	12.2851	11.9998	12.3145	12.3718
14/10/2015	GR9	6.8466	14.1066	14.2041	15.5851	15.4235	16.1579	16.3666
	HG1	12.1833	15.3207	15.3557	15.6182	15.5280	15.8507	16.0660
	HR1	3.9083	9.2177	9.2161	8.4127	8.2599	9.0035	9.2738
	HV1	5.3969	9.1737	9.2850	9.5684	9.5196	10.5241	10.6611
	KC1	5.5000	10.7263	10.8241	11.5256	11.5326	12.0574	12.3407
	KC2	16.3481	14.9944	14.8669	15.6873	15.7329	15.5961	15.6311
	LH0	4.9083	9.2240	9.2029	9.4828	9.8286	9.7004	9.9389
	LW2	12.0641	16.6002	17.0335	17.4637	17.5776	19.0537	19.4729
	ST5	11.7100	17.3403	17.2562	18.5583	18.6559	18.4578	18.6564
	TH4	11.4372	14.2911	14.1944	12.8099	12.5399	13.2005	13.0827
	RMSE		4.5915	4.6862	5.3756	5.3441	5.9838	6.1375
	BL0	9.7250	9.1394	9.5084	9.2373	9.0720	10.3669	10.3735
	BX1	8.7542	9.2452	9.6964	9.4575	9.4023	11.0635	10.9232
	CD1	15.7458	13.1982	13.8878	14.4752	14.6295	15.1654	15.2673
	CD9	16.7136	17.1385	17.6003	20.8312	20.8286	22.5164	22.4121
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	14.2139	11.9462	12.3860	12.7244	12.5511	13.3676	13.5510
	GN2	11.3665	13.0389	13.2809	13.0960	12.6332	14.2493	13.8839
	GN3	13.6809	14.5887	15.1096	16.2269	16.1551	17.2100	17.2348
	GR4	9.6750	9.7142	10.3442	11.6428	11.5418	12.5773	12.4138
15/10/2015	GR9	10.9183	12.5926	13.3553	14.4441	14.4793	15.7931	15.7575
	HG1	15.6625	14.2958	14.8705	14.8942	15.0120	15.8298	15.8158
	HR1	9.2792	8.7723	9.0586	7.8911	7.8735	9.0405	8.9754
	HV1	8.4433	9.0776	9.3153	9.9732	9.5705	11.2786	11.0771
	KC1	9.7583	9.6086	10.1795	10.7187	10.8735	11.7791	11.9142
	KC2	16.3766	14.8074	14.8209	15.5281	15.6666	15.5174	15.5794
	LH0	9.4458	8.4811	8.8371	9.0007	9.3475	9.2169	9.4995
	LW2	19.0288	16.1614	16.5295	17.9645	17.4071	18.4815	18.6262
	ST5	13.4300	16.5048	16.8378	18.0343	18.2470	18.2743	18.3947
	TH4	11.4372	14.3416	14.3138	13.3751	12.6875	13.5397	13.2460
	RMSE		1.6364	1.6407	2.0530	2.0327	2.6617	2.6014
	BL0	20.6917	8.7551	9.0939	8.4376	8.6889	8.9826	9.5777
	BX1	18.7625	8.8077	9.2328	8.5775	8.9259	9.8218	10.2111
	CD1	24.9708	12.6254	13.1037	13.2955	14.0773	13.9928	14.4236
	CD9	24.1276	16.8422	17.4575	19.7563	20.4992	20.9938	21.9770
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	21.1872	11.4863	11.7710	11.6293	12.0611	12.1606	12.4982
	GN2	22.0827	12.9648	13.1694	12.6335	12.4843	13.2021	13.2350
	GN3	24.6283	14.0634	14.4959	15.1209	15.6502	16.0770	16.4061
	GR4	19.5625	9.2150	9.9691	10.4229	11.0418	11.5530	11.8055
16/10/2015	GR9	20.7506	12.0283	12.5979	13.3237	13.8013	14.2595	14.6997
	HG1	25.6625	13.7161	14.2800	13.6095	14.5076	14.4002	15.0668
	HR1	18.1167	8.4186	8.7680	7.0971	7.5542	7.7154	8.4989
	HV1	18.6330	8.8334	9.1110	9.2829	9.3556	10.5497	10.7365
	KC1	19.4458	9.0852	9.5528	9.7070	10.3608	10.5054	11.0690
	KC2	24.0530	14.6171	14.6205	15.5388	15.6539	15.4924	15.5508
	LH0	17.0958	8.1215	8.3979	8.2185	9.0546	8.0409	8.9462
	LW2	26.6666	15.7789	16.0236	17.9847	17.1681	17.5616	17.5948
	ST5	24.2300	16.0325	16.2769	16.9967	17.8289	17.2384	17.7025
	TH4	11.4372	14.1349	14.1374	13.2759	12.5762	13.7454	13.0809
	RMSE		9.5196	9.1590	9.0333	8.6508	8.4599	8.0599

Table 5.5: Evaluation of applying the DSTM model to data from 01/03/15 to 13/10/15 considering multi-pollutant data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 14, 15 and 16 of October, 2015. All comparisons were made in the original scale (μg/m^3).

			7 Months		3 Months		2 Months	
	Site	Real	MM	EM	MM	EM	MM	EM
	BL0	6.3750	10.0900	10.3487	9.3043	9.2002	9.0709	9.1616
	BX1	5.2333	9.3845	9.4815	9.0730	8.6167	9.2244	9.3992
	CD1	10.8458	13.2643	13.4875	13.6205	13.6570	13.0414	13.2650
	CD9	10.6696	17.4862	17.6874	21.6403	21.1013	19.1652	18.6794
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	11.0350	13.7107	13.9098	14.1192	13.9760	11.8825	12.4382
	GN2	6.8770	13.6129	14.4297	13.7216	13.5957	13.9323	14.5182
	GN3	10.2575	14.8727	15.1599	15.5043	15.4311	15.2173	15.4904
	GR4	5.7458	10.4702	10.7741	11.2160	10.9060	10.7927	10.6867
14/10/15	GR9	6.8466	12.8707	13.1881	13.6542	13.3370	12.8903	12.9830
	HG1	12.1833	14.6309	14.7961	14.4228	14.4032	14.1681	14.1099
	HR1	3.9083	9.2528	9.1417	8.7384	8.6770	8.9031	8.7932
	HV1	5.3969	9.1951	9.2949	9.3093	9.1361	9.5148	9.3954
	KC1	5.5000	9.9322	10.1037	10.1137	10.0536	10.1439	10.3431
	KC2	16.3481	14.4665	14.6175	15.3162	15.4172	15.8117	15.8901
	LH0	4.9083	9.6856	10.2296	9.2819	9.5311	9.2123	9.4479
	LW2	12.0641	16.8520	17.6882	16.6663	17.0084	16.6207	18.4489
	ST5	11.7100	15.9916	16.2194	16.9794	17.1184	16.2362	16.3148
	TH4	11.4372	13.8150	13.2160	12.6097	12.1327	11.1190	11.0768
	RMSE		4.3389	4.5995	4.8503	4.7254	4.3821	4.5588
	BL0	9.7250	9.8828	10.3344	9.0593	9.0895	8.9940	9.0773
	BX1	8.7542	9.2751	9.4937	8.9163	8.5190	9.4956	9.4657
	CD1	15.7458	12.7770	13.2613	13.2281	13.2871	13.2872	13.3877
	CD9	16.7136	17.2393	17.7659	21.5417	21.2621	20.9221	19.1601
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	14.2139	12.7259	13.5429	12.8697	13.5005	11.5225	12.1952
	GN2	11.3665	14.0425	14.5395	13.9677	13.7006	15.9023	15.6137
	GN3	13.6809	14.5613	15.1028	15.2467	15.2384	15.2634	15.5829
	GR4	9.6750	10.2124	10.6989	11.0566	10.8369	10.4984	10.7371
15/10/15	GR9	10.9183	12.3374	12.8990	13.2332	13.0772	13.0479	13.2622
	HG1	15.6625	14.2764	14.6968	14.1790	14.2047	14.1931	14.2764
	HR1	9.2792	9.1274	9.1775	8.3354	8.3799	8.7630	8.7112
	HV1	8.4433	9.3414	9.5956	9.6685	9.4104	9.6841	9.6909
	KC1	9.7583	9.5235	9.9553	9.8397	9.8137	10.1902	10.2503
	KC2	16.3766	14.3615	14.6438	15.3088	15.3622	15.7573	15.8621
	LH0	9.4458	9.3489	10.1154	8.9451	9.2966	9.9588	9.6046
	LW2	19.0288	16.9061	17.6318	16.3989	16.4518	18.7155	18.8265
	ST5	13.4300	15.6930	16.1136	16.7331	16.8493	16.5659	16.6703
	TH4	11.4372	13.6688	13.3809	12.6353	12.2300	11.0820	11.1600
	RMSE		1.5100	1.5287	1.9690	1.8563	1.9899	1.7737
	BL0	20.6917	10.0017	10.2103	8.7779	8.8554	9.8026	9.0851
	BX1	18.7625	9.4069	9.3967	8.7374	8.2166	11.7599	9.5626
	CD1	24.9708	12.8431	12.9501	12.9905	12.8535	14.7610	13.5548
	CD9	24.1276	17.3643	17.7126	21.3564	21.1043	21.1395	19.3063
	CT3	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543	14.5543
	GN0	21.1872	12.3318	13.1483	12.6026	12.9568	13.5397	12.2420
	GN2	22.0827	14.2576	14.6161	13.9305	13.7156	17.8179	15.8129
	GN3	24.6283	14.6203	14.8917	15.0994	14.9015	17.2386	15.7027
	GR4	19.5625	10.2188	10.5062	10.9539	10.5337	12.3127	10.8652
16/10/15	GR9	20.7506	12.3804	12.5591	12.6543	12.6709	15.5560	13.6881
	HG1	25.6625	14.3413	14.4904	13.9632	13.8884	15.5599	14.4595
	HR1	18.1167	9.2132	9.1543	7.9897	8.0581	9.4928	8.7862
	HV1	18.6330	9.4756	9.6636	9.5868	9.4343	11.0077	9.8304
	KC1	19.4458	9.5702	9.7125	9.6607	9.4728	11.1354	10.2567
	KC2	24.0530	14.4321	14.5877	15.3265	15.3492	15.8055	15.8474
	LH0	17.0958	9.4548	9.9053	8.8308	9.0911	11.0469	9.6057
	LW2	26.6666	17.1217	17.3803	16.4847	16.0460	20.3997	18.6939
	ST5	24.2300	15.7538	15.8978	16.7506	16.5383	18.1778	16.9543
	TH4	11.4372	13.6354	13.3837	12.6169	12.1179	11.6128	11.2770
	RMSE		8.8834	8.6712	8.8459	8.9519	7.1664	8.3094

Table 5.6: Evaluation of applying the DSTM model to data from 01/03/15 to 13/10/15 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM) and expectation-maximisation (EM) approaches to inference (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 14, 15 and 16 of October, 2015. All comparisons were made in the original scale (μg/m^3).

Chapter 6

Spatial-Temporal Prediction

During the last years, there has been a increasing tendency to use statistical rather than just interpolation methods in air pollution modelling such as Kolehmainen *et al.* (2001). From a modern perspective, *geostatistics* is defined as a sub-branch of spatial statistics in which the data consists of a finite sample of measured values relating to an underlying spatially continuous phenomenon, Diggle & Ribeiro (2007). *Geostatistics* has its origins in the South African mining industry, but there was also an independent development for spatial prediction in the 1950s, (Mathéron, 1963). The spatial prediction method known as *kriging* is named in honour of D. G. Krige, who promoted the use of statistical methods in mineral exploration. Originally, geostatistical methods were developed to predict the likely yield of a mining operation over a spatial region, given the results of samples of minerals extracted from a finite set of locations.

Kriging has been at the centre stage of spatial statistics and recently, there has been strong interest in applying kriging to air pollution studies. There are several works describing theoretical geostatistics and some kriging applications on spatial interpolation using air pollution data, some references are Cressie (1993); Stein (1999); Janssena *et al.* (2008). There is an extension to spatio-temporal models that have also been used in in air pollution studies, see Carroll *et al.* (1997); Wikle *et al.* (1998); Kyriakidis & Journel (1999); Huerta *et al.* (2004) for further details. A Bayesian version of Kriging has also been suggested and has been applied to air pollution, for example see Le & Zidek (1992); Handcock & Stein (1993); Ecker & Gelfand (1997).

Kriging can be classified into three types: ordinary, simple and universal kriging, which differ in their assumptions about the structure of the mean: ordinary kriging assumes a constant, unknown mean; simple kriging assumes a constant, known mean; and

universal kriging assumes a trending mean. Additionally, data are often available for other variables at the sampled locations. If the secondary variables are correlated with the primary variable being modelled (and are known at the unsampled locations) they may be used to help predict conditions at unsampled locations by co-kriging (Oliver & Webster, 1990).

The references about spatio-temporal modelling are diverse. For example, a traditional covariance based model can be seen in Guttorp *et al.* (1994). In Tiao *et al.* (1975) an empirical model is proposed to explain the CO concentration and its trend. In recent years, hierarchical Bayesian approaches for spatial prediction of air pollution have been developed. The dynamic linear model framework offers an alternative to prediction, Wikle & Cressie (1999); Huerta *et al.* (2004); Shaddick & Wakefield (2002); Cressie & Wikle (2011).

One could be focused on prediction at some location in space within the time span of the observations and parameter inference for spatio-temporal covariates. If it is the case, then the true process can be written in terms of spatio-temporal fixed effects due to covariates plus a spatio-temporally dependent random process.

6.1 Spatio-temporal Kriging

Let

$$\{Y(\mathbf{s};t) : \mathbf{s} \in D_{\mathbf{s}}, t \in D_t\} \tag{6.1}$$

denote the spatio-temporal process that is a statistical model for a phenomenon. This hidden random process is defined for a spatial location $\mathbf{s} \in D_{\mathbf{s}}$ (a subset of d-dimensional Euclidean space), and time t in temporal domain D_t (along the one-dimensional real line).

Optimal predictions are required for a latent spatio-temporal process; this can be made through a linear spatial prediction (kriging). It can be formulated based on spatial statistical dependencies, such as a spatial covariance function or a spatial variogram.

Now, let us suppose the data in terms of the latent spatio-temporal process of interest plus a measurement error,

$$Z(\mathbf{s}_{ij};t_j) = Y(\mathbf{s}_{ij};t_j) + \epsilon(\mathbf{s}_{ij};t_j), \qquad i = 1,\ldots,m_j;\ \ j = 1,\ldots,T, \tag{6.2}$$

where the error $\{\epsilon(\mathbf{s}_{ij};t_j)\}$ is independent of $Y(\cdot;\cdot)$ and represent the measurement error that is assumed to be *iid* with mean zero and variance σ^2_ϵ.

In spatio-temporal setting, there are two different ways of writing the data vector $\mathbf{Z}$. It could be seen as a multivariate spatial process, where $\mathbf{Z}^{(i)} \equiv (Z(\mathbf{s}_i; t_{ij}) : j = 1, \ldots, T_i)'$; $i = 1, \ldots, m$, is the T_i-dimensional vector of temporal data. Alternatively, one could represent all data at time point t and call it $\mathbf{Z}_t$. The concatenation of such $\{\mathbf{Z}_t : t = 1, \ldots, T\}$ amounts to a reordering of the data vector $\mathbf{Z}$, that could also be written as $(\mathbf{Z}_1', \cdots, \mathbf{Z}_T')$. In the data model 6.2, it is assumed that the data are noisy observations of the latent process Y at a finite collection of locations in the space-time domain. Any of the two orderings of data in $\mathbf{Z}$ will be used, it will be determined by the parsimony of parameterisation.

In general, one has not necessarily observed data at all time points or locations. Sometimes it is assumed that data were observed at the same set of m locations for each of the T times, in which case $\mathbf{Z}$ is of length mT.

One would like to predict the hidden value $Y(\mathbf{s}_0; t_0)$, based on data defined by 6.2. Now suppose that the latent process follows the model

$$Y(\mathbf{s}; t) = \mu(\mathbf{s}; t) + \eta(\mathbf{s}; t) \tag{6.3}$$

for all $(\mathbf{s}, t)$ in our space-time domain of interest, such as $(D_\mathbf{s} \times D_t)$. In equation 6.3 $\mu(\mathbf{s}; t)$ denotes the process mean and $\eta(\mathbf{s}; t)$ denotes a mean-zero random process with spatial and temporal statistical dependence.

6.1.1 Prediction

Apart from characterise dependence in space and time, spatio-temporal covariances can be used to formulate an optimal linear spatial prediction. These predictors can be quantified through a spatial covariance or a spatial variogram.

For any unbiased linear predictor, $Y^*(\mathbf{s}_0; t_0)$, the mean squared prediction error,

$$E(Y^*(\mathbf{s}_0; t_0), Y(\mathbf{s}_0; t_0))^2, \qquad \text{for} \quad \mathbf{s}_0 \in D_\mathbf{s}, t_0 \in D_t, \tag{6.4}$$

can be expressed in terms of the covariance function. The goal of kriging is to find the optimal linear predictor from incomplete and noisy data. It is optimal in the sense that minimises the mean squared prediction error between $Y(\mathbf{s}_0, t_0)$ and our prediction, $\hat{Y}(\mathbf{s}_0, t_0)$. A prediction is sought at some spatio-temporal location $(\mathbf{s}_0, t_0)$. If $t_0 < t_T$, so that all is data available, then it is a smoothing situation. If data is known up to time t_0 then it is a filtering situation. And, if $t_0 > t_T$ then it is a forecasting situation.

In spatio-temporal context, recall from 4.20 that the *variogram* is defined as

$$2\gamma_Z(\mathbf{s}_i, \mathbf{s}_k; t_j, t_l) \equiv \text{var}(Z(\mathbf{s}_i; t_j) - Z(\mathbf{s}_k; t_l)), \quad \mathbf{s}_i, \mathbf{s}_k \in \mathbb{R}^d, t_j, t_l \in \mathbb{R}. \tag{6.5}$$

and from 4.1, recall the *spatio-temporal function*,

$$C(\mathbf{s}_i, \mathbf{s}_k; t_j, t_l) \equiv \text{cov}(Y(\mathbf{s}_i, t_j), Y(\mathbf{s}_k, t_l)). \tag{6.6}$$

There are different approaches to obtain the form of the optimal linear predictor, the spatio-temporal kriging. In one of these approaches, it is assumed that the underlying process is a *Gaussian process* and the measurement error has a Gaussian distribution.

A Gaussian process is denoted by $\{Y(\mathbf{r}) : \mathbf{r} \in D\}$, where $\mathbf{r}$ is a spatial, temporal, or spatio-temporal location in D, a subset of d-dimensional space. If the process has all its finite-dimensional distributions Gaussian, determined by a mean function $\mu(\mathbf{r})$ and a covariance function $C(\mathbf{r}, \mathbf{r}') = \text{cov}(Y(\mathbf{r}), Y(\mathbf{r}'))$, the Gaussian process is denoted by $Y(\mathbf{r}) \sim GP(\mu(\mathbf{r}), c(\cdot, \cdot))$ for any location $\{\mathbf{r}, \mathbf{r}'\} \in D$. In spatio-temporal statistics it is common to use $Gau(\cdot, \cdot)$ instead of $GP(\cdot, \cdot)$ and this convention will be followed.

The *simple kriging* predictor $\hat{Y}(\mathbf{s}_0, t_0)$, takes the form of a linear combination,

$$\begin{aligned} \hat{Y}(\mathbf{s}_0, t_0) &\equiv \sum_{i=1}^{m_j} \sum_{j=1}^{T} \boldsymbol{\ell}_{ij} \mathbf{Z}(\mathbf{s}_{ij}; t_j) + c \\ &\equiv \boldsymbol{\ell}' \mathbf{Z} + c, \end{aligned} \tag{6.7}$$

where $\mathbf{Z} \equiv (\mathbf{Z}_1', \ldots, \mathbf{Z}_T')'$, $\boldsymbol{\ell}$ and c are optimised. For the moment it is assumed that the mean function,

$$\mu(\mathbf{s}; t) = E(Y(\mathbf{s}; t)), \qquad \mathbf{s} \in D_{\mathbf{s}}, t \in D_t, \tag{6.8}$$

is known.

If it is assumed that the mean function is constant over space and time, then one can obtain the *ordinary-kriging* predictor. If the mean function is a linear combination of covariates, then one can derive the *universal-kriging* predictor.

In the context of Spatio-Temporal kriging, time is treated as another dimension, and the covariance functions should describe covariability between any two space-time locations. One should notice that time has a direction and different properties of causality than space. Thus, any proposed covariance functions should consider that duration in time

is different from distances in space.

The universal kriging can be illustrated, there are analogous ordinary-kriging and simple kriging equations that can be derived. Let us define the hidden process as

$$\mathbf{Y} \equiv (Y(\mathbf{s}_{11}; t_1), \cdots, Y(\mathbf{s}_{m_T T}; t_T))'$$

and the measurement error process as $\boldsymbol{\varepsilon} \equiv (\epsilon(\mathbf{s}_{11}; t_1), \cdots, \epsilon(\mathbf{s}_{m_T T}; t_T))'$. The data model can be expressed in terms of vectors,

$$\mathbf{Z} = \mathbf{Y} + \boldsymbol{\varepsilon}. \tag{6.9}$$

In a similar way, the vector form of the process model for Y is written

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\eta}, \tag{6.10}$$

where $\boldsymbol{\mu} \equiv (\mu(\mathbf{s}_{11}; t_1), \cdots, \mu(\mathbf{s}_{m_T T}; t_T))' = \mathbf{X}\boldsymbol{\beta}$, and $\boldsymbol{\eta} \equiv (\eta(\mathbf{s}_{11}; t_1), \cdots, \eta(\mathbf{s}_{m_T T}; t_T))'$.

Note that $\text{cov}(\mathbf{Y}) \equiv \mathbf{C}_Y = \mathbf{C}_\eta$, $\text{cov}(\boldsymbol{\varepsilon}) \equiv \mathbf{C}_\epsilon$ and $\text{cov}(\mathbf{Z}) \equiv \mathbf{C}_Z = \mathbf{C}_Y + \mathbf{C}_\epsilon$. Now, defining $\mathbf{c}_0' \equiv \text{cov}(Y(\mathbf{s}_0; t_0), \mathbf{Z})$, $c_{0,0} = \text{var}(Y(\mathbf{s}_0; t_0))$ and $\mathbf{X}$ the $(\sum_{j=1}^{T} m_j) \times p$ matrix given by $\mathbf{X} \equiv [\mathbf{x}(\mathbf{s}_{ij}; t_j)' : i = 1, \ldots, m_j; j = 1, \ldots, T]$, consider the joint Gaussian distribution,

$$\begin{bmatrix} Y(\mathbf{s}_0; t_0) \\ \mathbf{Z} \end{bmatrix} \sim Gau\left(\begin{bmatrix} \mathbf{x}(\mathbf{s}_0; t_0)' \\ \mathbf{X} \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} c_{0,0} & \mathbf{c}_0' \\ \mathbf{c}_0 & \mathbf{C}_Z \end{bmatrix} \right). \tag{6.11}$$

Recall the simple kriging and for now assume that $\boldsymbol{\beta}$ is known, the conditional distribution can be obtained,

$$Y(\mathbf{s}_0; t_0)|\mathbf{Z} \sim Gau(\mathbf{x}(\mathbf{s}_0; t_0)'\boldsymbol{\beta} + \mathbf{c}_0'\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), c_{0,0} - \mathbf{c}_0'\mathbf{C}_Z^{-1}\mathbf{c}_0). \tag{6.12}$$

Under the Gaussian assumption 6.11, the spatio-temporal universal kriging predictor is the posterior mean of 6.12, i.e.,

$$\hat{\mathbf{Y}}(\mathbf{s}_0; t_0) = \mathbf{x}(\mathbf{s}_0; t_0)'\boldsymbol{\beta} + \mathbf{c}_0'\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}). \tag{6.13}$$

The variance is the spatio-temporal simple kriging variance,

$$\sigma^2_{Y,\text{sk}}(\mathbf{s}_0; t_0) = c_{0,0} - \mathbf{c}_0'\mathbf{C}_Z^{-1}\mathbf{c}_0. \tag{6.14}$$

Note that $\sigma_{Y,\text{sk}}(\mathbf{s}_0; t_0)$ is called the spatio-temporal simple kriging *prediction standard*

error and it has the same units as $\hat{Y}(\mathbf{s}_0; t_0)$.

If one considers the spatio-temporal kriging from the Gaussian-process perspective, then a valid finite-dimensional Gaussian distribution for any finite subset of locations can be specified. For this purpose, the covariance between the process at any two locations in the domain of interest, $\mathbf{c}_0$, should be specified.

Note that the weights $\mathbf{w}' \equiv \mathbf{c}_0' \mathbf{C}_Z^{-1}$ are a function of the covariances and the measurement-error variance. The conditional mean in 6.13 takes the residuals between the observations and their marginal means, $(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})$ with their respective weights $\mathbf{w}$.

In most of the cases, one would not know $\boldsymbol{\beta}$, in particular the optimal linear unbiased predictor, or spatio-temporal *universal kriging* (UK) predictor of $\mathbf{Y}(\mathbf{s}_0; t_0)$ is

$$\hat{\mathbf{Y}}(\mathbf{s}_0; t_0) = \mathbf{x}(\mathbf{s}_0; t_0)' \hat{\boldsymbol{\beta}}_{\text{gls}} + \mathbf{c}_0' \mathbf{C}_Z^{-1} (\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{gls}}), \tag{6.15}$$

where the generalised least squares (gls) estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\text{gls}} \equiv (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{Z}. \tag{6.16}$$

The associated spatio-temporal universal kriging variance is given by

$$\sigma^2_{Y,\text{uk}} = c_{0,0} - \mathbf{c}_0' \mathbf{C}_Z^{-1} \mathbf{c}_0 + \kappa, \tag{6.17}$$

where

$$\kappa \equiv (\mathbf{x}(\mathbf{s}_0; t_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_0)' (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X}) (\mathbf{x}(\mathbf{s}_0; t_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_0) \tag{6.18}$$

is the additional uncertainty brought to the prediction due to the estimation of $\boldsymbol{\beta}$. The expression $\sigma_{Y,\text{uk}}$ is called the spatio-temporal universal kriging prediction standard error.

In general, one rarely knows the variances and covariances that are involved, $\mathbf{C}_Y$, $\mathbf{C}_\epsilon$, $\mathbf{c}_0$ and $c_{0,0}$. One possible solution is to parameterise them, and then estimate these parameters through maximum likelihood, restricted maximum likelihood or through a fully Bayesian approach.

In the next section two models will be applied based on two different covariance functions. The first of these covariance functions is given by an isotropic and stationary separable model of the form given by 4.10,

$$C^{(\text{sep})}(\mathbf{h}; \tau) = C^{(\mathbf{s})}(\mathbf{h}) \cdot C^{(t)}(\tau), \qquad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}. \tag{6.19}$$

Both covariance functions, $C^{(\mathbf{s})}(\cdot)$ and $C^{(t)}(\cdot)$, take the form

$$C^{(\cdot)} = b_1 \exp(-\phi h) + b_2 I(h = 0), \tag{6.20}$$

where ϕ, b_1 and b_2 are parameters that are different for $C^{(\mathbf{s})}(\cdot)$ and $C^{(t)}(\cdot)$. All parameters need to be estimated and $I(\cdot)$ is the indicator function that is used to represent the *nugget effect*.

The second fitted model is a non-separable spatio-temporal covariance function. The temporal lag is scaled to account for the different nature of space and time. The model is given by

$$C^{(\text{st})}(||\mathbf{v}_a||) \equiv b_1 \exp(-\phi||\mathbf{v}_a||) + b_2 I(||\mathbf{v}_a|| = 0), \tag{6.21}$$

where $\mathbf{v}_a \equiv (\mathbf{h}' a\tau)'$, and recall that $||\mathbf{v}_a|| = (\mathbf{h}'\mathbf{h} + a^2\tau^2)^{1/2}$. In this case a is the scaling factor used for generating the space-time anisotropy. In figure 6-3, fitted separable and non-separable semivariogram for $\log(\text{PM}_{2.5})$ levels during the period of interest are shown.

The first semivariogram considered here corresponds to the spatio-temporal separable covariance function in 4.10 and 6.20. The separable covariance function 4.10 corresponds to a semivariogram of the form

$$\gamma^{\text{sep}}(\mathbf{h}; \tau) = \text{sill} \cdot \left(\bar{\gamma}^{(\mathbf{s})}(||\mathbf{h}||) + \bar{\gamma}^{(t)}(|\tau|) - \bar{\gamma}^{(\mathbf{s})}(||\mathbf{h}||)\bar{\gamma}^{(t)}(|\tau|) \right), \tag{6.22}$$

where the standardised semivariograms $\bar{\gamma}^{(\mathbf{s})}$ and $\bar{\gamma}^{(t)}$ have separate nugget effects and sills equal to 1. For more detalis see Gräler *et al.* (2016); Montero *et al.* (2015).

6.1.2 Likelihood Estimation

Given the data model 6.9, recall that $\mathbf{C}_Z = \mathbf{C}_Y + \mathbf{C}_\epsilon$. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_Y, \boldsymbol{\theta}_\epsilon\}$ denote the parameters for the covariance functions of the hidden process Y and the measurement-error process ϵ respectively. With an abuse of notation, the likelihood can then be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{Z}) \propto |\mathbf{C}_z(\boldsymbol{\theta})|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{C}_Z(\boldsymbol{\theta}))^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \right\}. \tag{6.23}$$

The aim is to maximise this likelihood function with respect to $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$ to obtain the maximum likelihood estimates (MLEs), $\{\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\boldsymbol{\theta}}_{\text{mle}}\}$.

Numerical methods are needed for maximisation of 6.23, because for most parametric covariance models, an analytical solution is not possible. Observe that covariance

parameters appear in the matrix inverse and determinant in likelihood.

To reduce the number of parameters in maximisation, the *profile likelihood* can be considered. For this purpose, the parameter $\boldsymbol{\beta}$ is replaced in 6.23 with the generalised least squares estimator,

$$\boldsymbol{\beta}_{\text{gls}} = (\mathbf{X}'\mathbf{C}_Z^{-1}(\boldsymbol{\theta})\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}_Z^{-1}(\boldsymbol{\theta})\mathbf{Z}. \tag{6.24}$$

Then, the profile likelihood is a function of the unknown parameters $\boldsymbol{\theta}$ that can be optimised through a numerical method to obtain $\hat{\boldsymbol{\theta}}_{\text{mle}}$. Now, the MLE of $\boldsymbol{\beta}$ can be obtained,

$$\hat{\boldsymbol{\beta}}_{\text{mle}} = (\mathbf{X}'\mathbf{C}_Z^{-1}(\hat{\boldsymbol{\theta}}_{\text{mle}})\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}_Z^{-1}(\hat{\boldsymbol{\theta}}_{\text{mle}})\mathbf{Z}, \tag{6.25}$$

Finally, the parameter estimates $\{\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\boldsymbol{\theta}}_{\text{mle}}\}$ are then substituted into the kriging equations 6.15 and 6.17. The results are the empirical best linear unbiased predictor (EBLUP) and the associated empirical prediction variance respectively.

An alternative to profiling is the *restricted maximum likelihood* (REML). Let us consider a contrast matrix $\mathbf{K}$ such that $E(\mathbf{KZ}) = \mathbf{0}$. A candidate can be an $(m - p - 1) \times m$ matrix orthogonal to the column space of the $m \times p$ design matrix $\mathbf{X}$, i.e., $\mathbf{K}$ corresponds to the $(m - p - 1)$ linearly independent rows of $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})$.

Since $\mathbf{KX} = 0$, then $E(\mathbf{KZ}) = \mathbf{KX}\boldsymbol{\beta} = \mathbf{0}$, and $\text{var}(\mathbf{KZ}) = \mathbf{KC}_Z(\boldsymbol{\theta})\mathbf{K}')$. Then, the estimate $\hat{\boldsymbol{\theta}}_{\text{reml}}$ can be obtained using numerical maximisation from the restricted likelihood,

$$L_{\text{reml}}(\boldsymbol{\theta}; \mathbf{Z}) \propto |\mathbf{KC}_Z(\boldsymbol{\theta})\mathbf{K}')|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{KZ})'(\mathbf{KC}_Z(\boldsymbol{\theta})\mathbf{K}')^{-1}(\mathbf{KZ})\right\}. \tag{6.26}$$

Note that parameter estimation and statistical inference with REML do not depend on the specific choice of $\mathbf{K}$, so long as it leads to $E(\mathbf{KZ}) = 0$. The GLS formula for $\boldsymbol{\beta}$, can be used to obtain

$$\hat{\boldsymbol{\beta}}_{\text{rmle}} = (\mathbf{X}'\mathbf{C}_Z^{-1}(\hat{\boldsymbol{\theta}}_{\text{rmle}})\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}_Z^{-1}(\hat{\boldsymbol{\theta}}_{\text{rmle}})\mathbf{Z}. \tag{6.27}$$

Finally, the parameter estimates $\{\hat{\boldsymbol{\beta}}_{\text{rmle}}, \hat{\boldsymbol{\theta}}_{\text{rmle}}\}$ are substituted into the kriging equations 6.15 and 6.17.

6.2 Implementation using the LAQN data

In universal kriging, the main purpose is to find the optimal linear predictor in the sense that it minimises the mean squared prediction error between $Y(\mathbf{s}_0; t_0)$ and our prediction, $\hat{Y}(\mathbf{s}_0; t_0)$. Depending on the situation, one can choose to let $\mu(\mathbf{s}; t)$ be:

- known,
- constant but unknown,
- modelled in terms of covariates, $\mu(\mathbf{s}; t) = \mathbf{x}(\mathbf{s}; t)'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is unknown.

These choices result in spatio-temporal models: simple, ordinary and universal kriging respectively previously discussed in section 6.1.2 .

6.2.1 Evaluation strategy

In order to assess the accuracy of the different Kriging approaches in producing spatial predictions of levels of $PM_{2.5}$, we perform a series of studies and evaluate how well they perform, based upon the RMSE. The basis of the evaluations is similar to that presented in 5.10.1 and here we use data on $PM_{2.5}$ levels from 01 August 2015 to 14 March 2016 and predict for 15, 16 and 17 March 2016. Three spatial locations are omitted from the data sets used to fit the model (sites BL0, GN3 and ST5) and then (out-of-sample) predictions three-days-ahead in both space and time are calculated for the three locations. Results are given for Kriging for $PM_{2.5}$, PM_{10} and NO_2 (Section 6.2.2) and for Fixed Rank Kriging (FRK) for $PM_{2.5}$ (Section 6.4). This is designed to provide a computationally efficient approach to dealing with large spatio-temporal covariance matrices.

6.2.2 Results

Two analogous analyses are also given for PM_{10} and NO_2. Their empirical spatio-temporal semivariogram of daily data set from 01 August 2015 to 15 March 2016 can be seen in figure 6-1 and figure 6-2 respectively. The empirical spatio-temporal semivariogram of daily PM_{25} data was previously presented in figure 4-6. The empirical semivariogram can be seen as a measure of dissimilarity in space/time. The lower the semivariogram value, the higher the correlation between two pairs of data points. In figures 6-1 and 6-2, one can note that a temporal correlation up to 5 days apart is observed and a strong spatial correlation up to 0.6 Km is present. Furthermore, it is clear that the correlations are higher in the PM_{10} data than NO_2. In addition, a clear

nugget effect is observed in both cases. Finally, one can conclude that there is some evidence of spatial-temporal interaction.

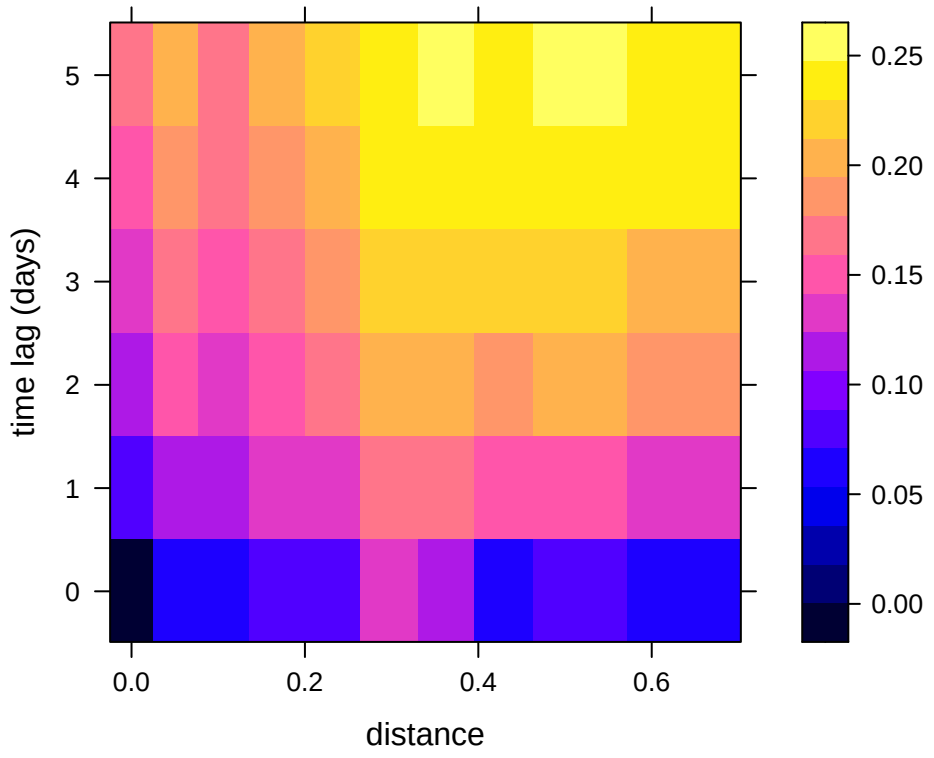

Figure 6-1: Empirical spatio-temporal semivariogram of daily PM_{10} data set from 01 August 2015 to 15 March 2016. The plot is produced with cutoff=1Km, bin width=0.1Km and time lags from 0 to 5 days.

The form of equation 6.20 assumes an exponential relationship between correlation and distance, with correlation decreasing with greater distances. This is a commonly used form, although other functions are often used, as shown in figure 4-1. One can compare the fits of different correlation-distance relationships, i.e. different semivariograms, by checking the mean squared error (MSE) of the fits. In this case, there was little difference between the MSEs found with the different models shown in figure 4-1 and so due to its simplicity and common use and as it proved a reasonable fit in comparison to other functions, the exponential model was used in the following analyses.

The non-separable spatio-temporal covariance function 6.21 can be used for space-time anisotropy, but it is relatively inflexible. It only contains four parameters, one

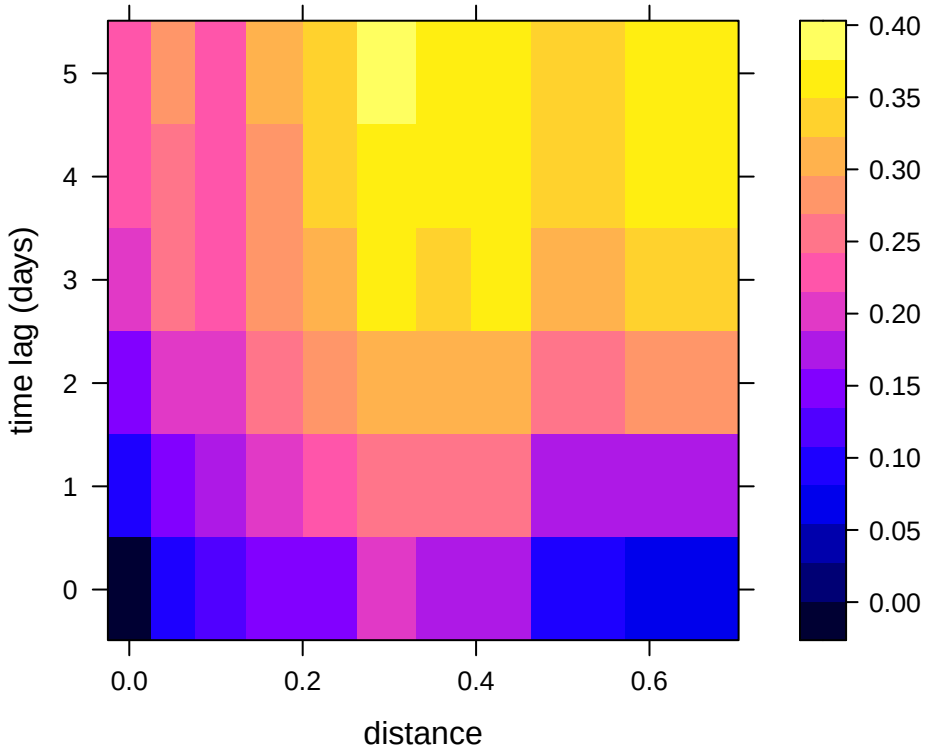

Figure 6-2: Empirical spatio-temporal semivariogram of daily NO_2 data set from 01 August 2015 to 15 March 2016. The plot is produced with cutoff=1Km, bin width=0.1Km and time lags from 0 to 5 days.

parameter (a) to account for the different scaling needed for space and time. One more parameter (ϕ) for the length scale, and two parameters to specify the variance (b_1, b_2). In contrast, the model 4.11 contains six parameters that provide a better reconstruction of the empirical covariance function. Although the separable model fits better in this case (in terms of the MSE), it is still a rather unrealistic model for most processes of interest.

It is important to notice that pollutants such as $PM_{2.5}$ are complex combinations of particles. Therefore, assumptions for most of the models that attempt to study these type of pollutants can be restrictive. For instance, assuming space-time anisotropy into a model to analyse $PM_{2.5}$ is unrealistic. However, some of these assumptions can be useful and convenient in producing an accurate forecast. One should try to find a balance between making reasonable assumptions, getting useful results and being

pragmatic.

Sites out: BL0, GN3, ST5				
Model	Site	RE: 15/03/2016	RE: 16/03/2016	RE: 17/03/2016
	BL0	2.86%	7.21%	24.74%
Separable	GN3	15.37%	10.49%	15.65%
	ST5	7.65%	9.34%	8.32%
RMSE		2.0812	1.6352	5.5828
	BL0	8.20%	6.13%	27.25%
Metric	GN3	13.83%	9.20%	18.97%
	ST5	16.65%	10.81%	16.58%
RSME		2.7179	1.6039	6.6713

Table 6.1: Evaluation of applying the kriging model to data from 01/08/15 to 17/03/16 considering $PM_{2.5}$ data. Results are shown using 16 sites from London data when fitting the model, using the separable and metric approach (see text for details). The relative error (RE) per site and date (in percentage) and the overall root mean squared error (RMSE) are shown. Evaluations are presented for 15, 16 and 17 of March 2016. The comparison was made in the original scale ($\mu g/m^3$).

Sites out: BL0, GN3, ST5				
Model	Site	RE: 15/03/2016	RE: 16/03/2016	RE: 17/03/2016
	BL0	4.02%	0.17%	9.40%
Separable	GN3	9.97%	10.07%	8.75%
	ST5	12.98%	1.06%	3.34%
RMSE		1.4751	2.2029	3.1561
	BL0	3.90%	6.89%	2.51%
Metric	GN3	5.36%	14.83%	5.01%
	ST5	18.18%	2.67%	6.71%
RSME		1.2349	2.1742	4.5518

Table 6.2: Evaluation of applying the kriging model to data from 01/08/15 to 17/03/16 considering PM_{10} data. Results are shown using 16 sites from London data when fitting the model, using the separable and metric approach (see text for details). The relative error (RE) per site and date (in percentage) and the overall root mean squared error (RMSE) are shown. Evaluations are presented for 15, 16 and 17 of March 2016. The comparison was made in the original scale ($\mu g/m^3$).

Sites out: BL0, GN3, ST5				
Model	Site	RE: 15/03/2016	RE: 16/03/2016	RE: 17/03/2016
	BL0	0.45%	8.04%	47.00%
Separable	GN3	7.37%	22.76%	119.73%
	ST5	31.07%	0.68%	70.24%
RMSE		8.9131	12.0440	20.2861
	BL0	3.90%	6.89%	2.51%
Metric	GN3	5.36%	14.83%	5.01%
	ST5	18.18%	2.67%	6.71%
RSME		1.2349	2.1742	4.5518

Table 6.3: Evaluation of applying the kriging model to data from 01/08/15 to 17/03/16 considering NO_2 data. Results are shown using 16 sites from London data when fitting the model, using the separable and metric approach (see text for details). The relative error (RE) per site and date (in percentage) and the overall root mean squared error (RMSE) are shown. Evaluations are presented for 15, 16 and 17 of March 2016. The comparison was made in the original scale ($\mu g/m^3$).

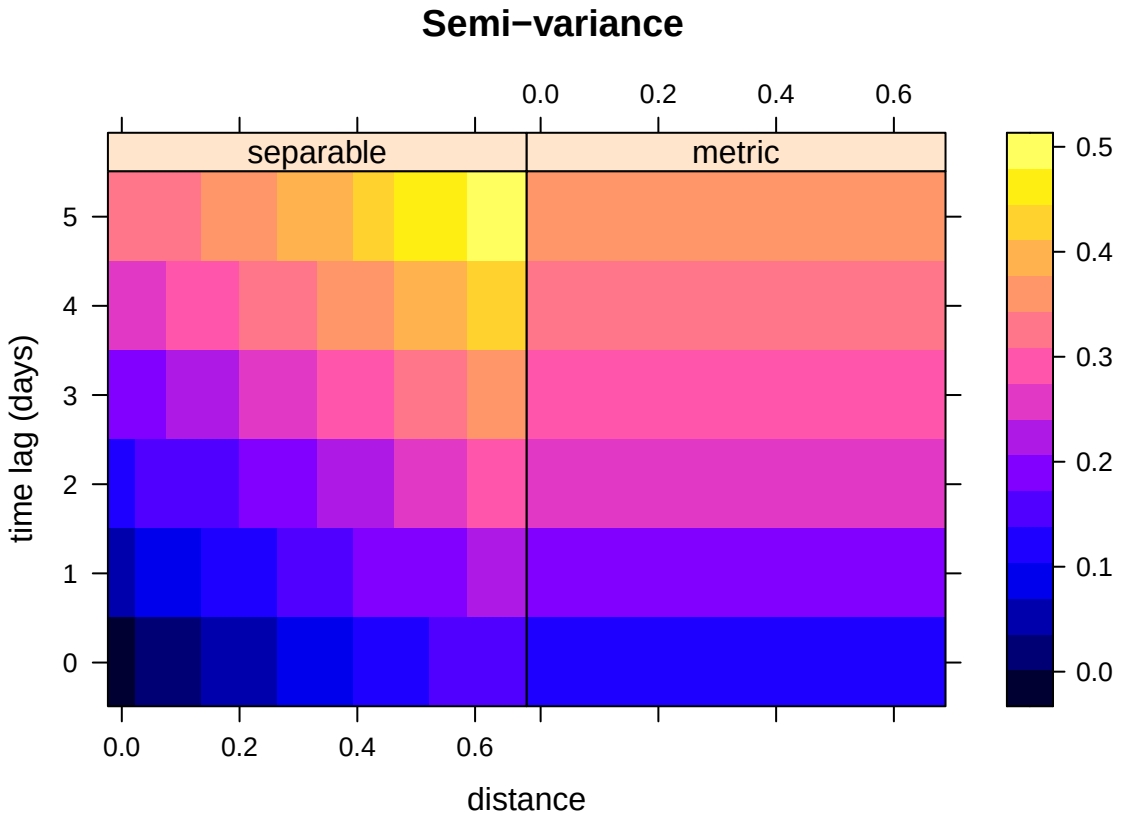

Figure 6-3: Fitted separable (left) and non-separable (right) semivariogram for $\log(PM_{2.5})$ levels from 01 August 2015 to 14 March 2016.

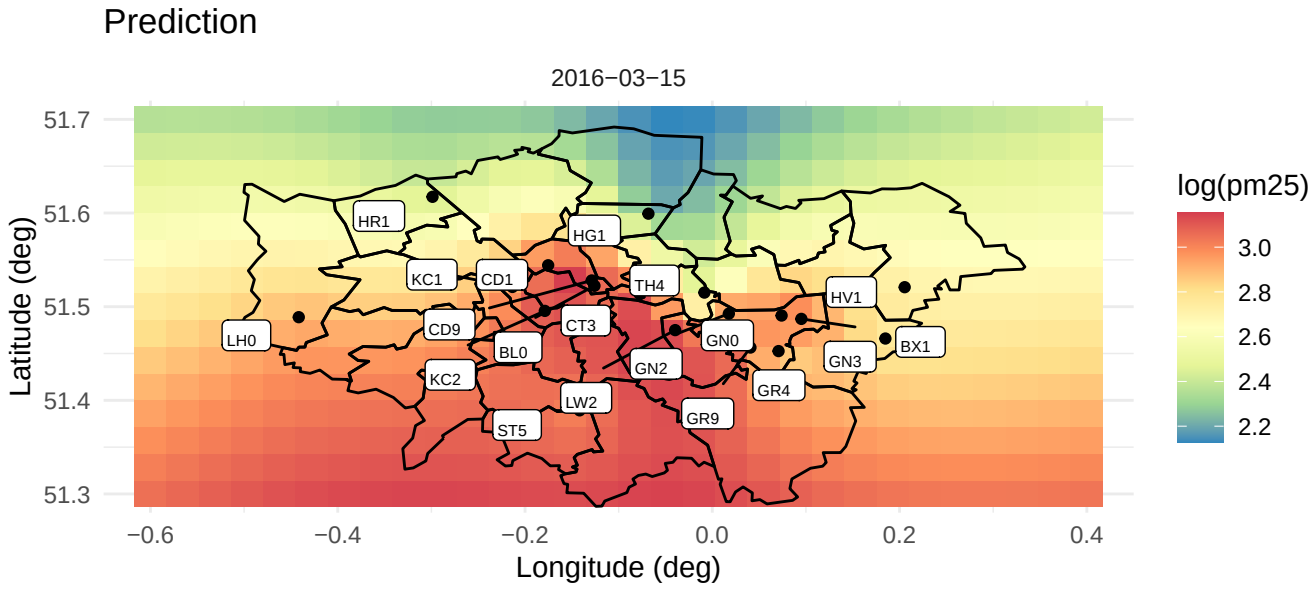

Figure 6-4: Spatio-temporal universal kriging predictions of $\log(\mathrm{PM}_{2.5})$ within a square latitude-longitude box enclosing the domain of interest 15 March 2016.

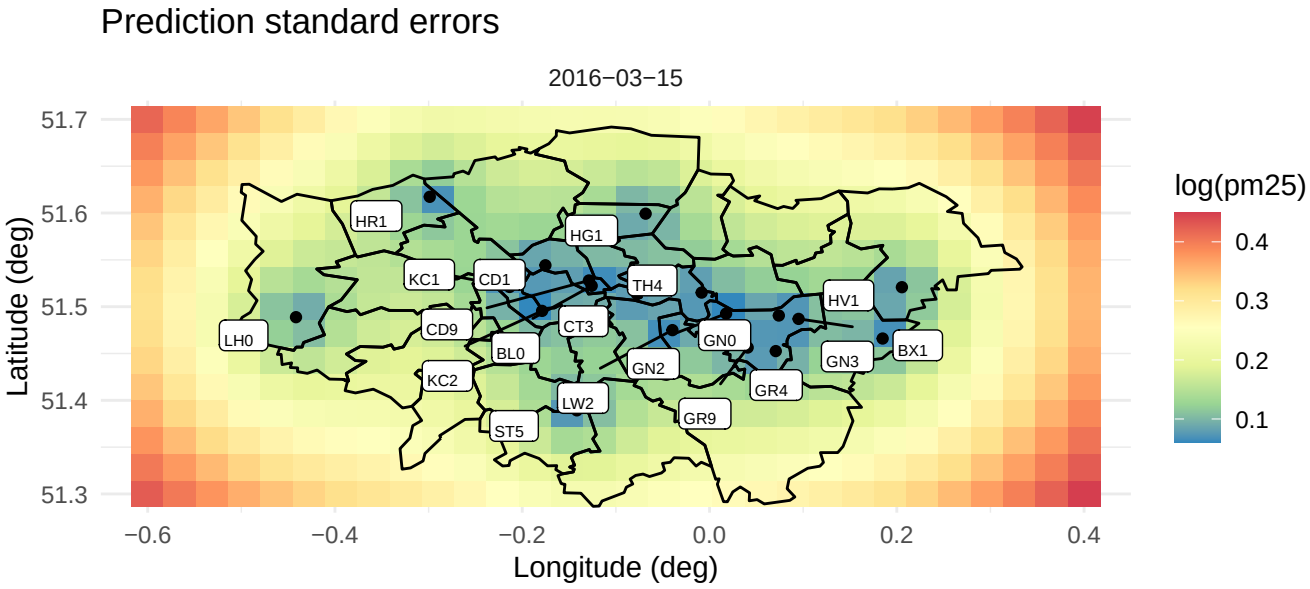

Figure 6-5: Spatio-temporal universal kriging prediction standard errors of $\log(\mathrm{PM}_{2.5})$ within a square latitude-longitude box enclosing the domain of interest 15 March 2016.

Both the spatio-temporal simple and universal kriging equations shown in section 6.1.2 can be extended to accommodate prediction at many locations in space and time. The predictions for $\log(PM_{2.5})$ 15 of March 2016 are shown in Figure 6-4. The model has been fitted from 1 August 2015 to 15 March 2016 on a space-time grid, considering a separable spatio-temporal covariance function. This plot can be compared with the results obtained from the IWD model, figure 2-9.

The corresponding prediction standard errors for the spatio-temporal universal kriging prediction are shown in figure 6-5. The errors associated with those for 15 March are considerably large, specially in areas close to the edges. Spatio-temporal kriging as shown in this example is relatively easy to implement for small data sets. However it starts to become prohibitive as data sets grow in size, unless some approximation is used. For example, the package `gstat` allows one to use an option to determine the maximum number of observations to use when doing prediction. The predictor is no longer optimal, but it is close enough to the optimal predictor for practical purposes.

A study was made in order to evaluate the spatial predictions for $PM_{2.5}$ obtained from the separable and metric models that were defined above. The idea is fitting the model considering data from 01 August 2015 to 17 March 2016. Instead of using 19 sites from LAQN, three sites are leaved out from the fitting (BL0, GN3, ST5). The same settings were considering for this exercise, the only difference is the number of analysed sites and two more days were included (16 and 17 March). As one can observe in table 6.1, the best model in terms of RMSE for prediction is the separable model. This comparison was made in the original scale. Different initial values were supplied to the optimisation routine used for fitting, but no differences were observed.

Finally, two analogous studies were done for PM_{10} and NO_2, see the results in table 6.2 and table 6.3 respectively. For PM_{10} data, the best model in terms of RMSE for prediction is the metric model. However the results are similar to the separable approach. On the other hand, the metric model for NO_2 data shows a better performance in terms of this analysis.

6.3 Random effects

In the context of spatial or spatio-temporal modelling, one of the most difficult problems is to specify realistic valid spatio-temporal covariance functions and to work with large spatio-temporal covariance matrices. If one assumes the model defined in 6.9, it is possible to parameterise $\mathbf{C}_Z$ in the context of large numbers of prediction or observation locations. One way to tackle these situations is take advantage of conditional

specifications that the hierarchical modelling framework allows.

There are different approaches for classical linear mixed models. From a marginal perspective, the random effects can be integrated and then, the resulting marginal distribution can be modelled. Consider the conditional representation of a general linear mixed-effects model for response vector $\mathbf{Z}$ and fixed and random effects vectors, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively,

$$\begin{aligned} \mathbf{Z}|\boldsymbol{\alpha} &\sim Gau(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha}, \mathbf{C}_\epsilon) \\ \boldsymbol{\alpha} &\sim Gau(\mathbf{0}, \mathbf{C}_{\boldsymbol{\alpha}}), \end{aligned} \tag{6.28}$$

where $\mathbf{X}$ and $\boldsymbol{\Phi}$ are assumed to be known matrices, and $\mathbf{C}_\epsilon$ and $\mathbf{C}_\alpha$ are known covariance matrices. Recall $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\epsilon, \boldsymbol{\theta}_\alpha\}$ are the covariance parameters in $\mathbf{C}_Z$ and $\mathbf{C}_\alpha$. For now, $\boldsymbol{\theta}$ can be omitted but it is implicitly considered.

The marginal distribution of $\mathbf{Z}$ is then given by integrating out the random effects,

$$[\mathbf{Z}|\boldsymbol{\theta}] = \int [\mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\theta}][\boldsymbol{\alpha}|\boldsymbol{\theta}] d\boldsymbol{\alpha}. \tag{6.29}$$

This distribution can be obtained by making use of iterated conditional expectation and variance formulas. The model associated with 6.29 can be written as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim Gau(\mathbf{0}, \mathbf{C}_\epsilon), \tag{6.30}$$

and

$$E(\mathbf{Z}) = E_\alpha\{E(\mathbf{Z}|\boldsymbol{\alpha})\} = E_\alpha\{\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha}\} = \mathbf{X}\boldsymbol{\beta} \tag{6.31}$$

$$\text{var}(\mathbf{Z}) = \text{var}_\alpha\{E(\mathbf{Z}|\boldsymbol{\alpha})\} + E_\alpha\{\text{var}(Z|\boldsymbol{\alpha})\} = \boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\epsilon. \tag{6.32}$$

Since 6.30 shows that $\mathbf{Z}$ is a linear combination of normally distributed random variables, it is normally distributed and the marginal distribution is given by

$$\mathbf{Z} \sim Gau(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\epsilon.) \tag{6.33}$$

One can observe that the conditional covariance matrix in 6.28, $\mathbf{C}_\epsilon$, is simpler compared to error covariance structure in the marginal distribution in 6.33, $\boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\epsilon$. Conditioning on random effects and inducing dependence through integration is important to hierarchical statistical modelling. The idea is to model means instead of to model covariances. The modelling effort should be put into the conditional mean, the integration induces a more complicated marginal dependence rather than specifying it

directly.

6.3.1 Basis-Function Representations

Covariance specification and the inconvenience of high dimensionality are the most difficult problems in spatio-temporal statistics to deal with. it is also needed to pay attention the choice of $\mathbf{\Phi}$ in 6.33, one option to tackle these problems is through basis-function expansions.

If it is supposed that one has a complex curve or surface in space. It is possible to decompose this curve or surface as a linear combination of some "elemental" basis functions. In this context, one can consider spatio-temporal basis functions to reconstruct the observed data.

Basis functions can be nonlinear functions of $(\mathbf{s}, t)$ and the associated coefficients can be inferred in a statistical additive model framework. There are different types of coefficients. If they are fixed but unknown, then one has a regression model and the basis functions can be thought as covariates. If the coefficients are random, then one has a random-effects model or, if covariates are also considered, it is a mixed-effects model. This framework allow us to build more complex models to be analysed through marginalisation.

6.3.2 Random Effects with Spatio-Temporal Basis Functions

Let us assume the same data model 6.9 and the process model defined in 6.3. It can be rewritten in terms of fixed effects $\boldsymbol{\beta}$, and random effects $\{\alpha_i : 1, \ldots, n_\alpha\}$,

$$Y(\mathbf{s}; t) = \mathbf{x}(\mathbf{s}; t)'\boldsymbol{\beta} + \eta(\mathbf{s}; t) = \mathbf{x}(\mathbf{s}; t)'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha} \phi_i(\mathbf{s}; t)\alpha_i + \nu(\mathbf{s}; t), \tag{6.34}$$

where $\{\phi_i(\mathbf{s}; t) : 1, \ldots, n_\alpha\}$ are spatio-temporal basis functions corresponding to location $(\mathbf{s}; t)$. Finally, the random effects are denoted by $\{\alpha_i\}$, and $\nu(\mathbf{s}; t)$ could represent small-scale spatio-temporal random effects not captured by the basis functions.

Let $\boldsymbol{\alpha} \sim Gau(\mathbf{0}, \mathbf{C}_\alpha)$, where $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_{n_\alpha})'$. Consider the process Y at n_y spatio-temporal locations, denoted by the vector $\mathbf{Y}$. The process can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{\Phi}\boldsymbol{\alpha} + \boldsymbol{\nu}, \tag{6.35}$$

where the ith column of the $n_y \times n_\alpha$ matrix $\mathbf{\Phi}$ corresponds to the ith basis function, $\phi(\cdot, \cdot)$, at all of the n_y spatio-temporal locations, in the same order given by in

$\mathbf{Y}$. The spatio-temporal ordering for the vector $\boldsymbol{\nu}$ is the same as the one in $\mathbf{Y}$, and $\boldsymbol{\nu} \sim Gau(\mathbf{0}, \mathbf{C}_\nu)$. Under this assumptions, the marginal distribution of $\mathbf{Y}$ is given by $\mathbf{Y} \sim Gau(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\nu)$, thus $\mathbf{C}_Y = \boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\nu$. The vector of covariance parameters $\boldsymbol{\theta}$ should include parameters in $\mathbf{C}_{\boldsymbol{\nu}}$. The spatio-temporal basis functions, $\boldsymbol{\Phi}$, contains the spatio-temporal dependence. In general, this relation can be non-separable. Observe that the random effects $\boldsymbol{\alpha}$ are not indexed by space and time, so it should be easier to specify a model for them. In this case, a model can be specified through a covariance matrix, which is easier than specifying a covariance function.

In a low-rank representation (when $n_\alpha \ll n_y$) under the model 6.35, one should note that $\mathbf{C}_z = \boldsymbol{\Phi}\mathbf{C}\boldsymbol{\Phi}' + \mathbf{V}$, where $\mathbf{V} \equiv \mathbf{C}_\nu + \mathbf{C}_\epsilon$. Then, considering the Sherman–Morrison–Woodbury matrix identity,

$$\mathbf{C}_z^{-1} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{V}^{-1}\boldsymbol{\Phi} + \mathbf{C}_\alpha^{-1})^{-1}\boldsymbol{\Phi}'\mathbf{V}^{-1}. \tag{6.36}$$

If $\mathbf{V}^{-1}$ is sparse or diagonal and $n_\alpha \ll n_y$, then this inverse is a function of a simple high-dimensional matrix $\mathbf{V}^{-1}$ and a low-dimensional matrix inverse $\mathbf{C}_\alpha^{-1}$. In the full rank ($n_\alpha = n_y$) and over-complete ($n_\alpha > n_y$) cases there are computational benefits through induced sparsity in $\mathbf{C}_\alpha$. Some basis-function implementations can assume that $\boldsymbol{\nu}$ and that $\boldsymbol{\Phi}$ is orthogonal, then $\boldsymbol{\Phi}\boldsymbol{\Phi}' = \mathbf{I}$. In these cases, one can reduce the computational burden significantly. In the product $\boldsymbol{\Phi}\boldsymbol{\alpha}$, its coefficients $\boldsymbol{\alpha}$, are random and the columns of $\boldsymbol{\Phi}$ are spatio-temporally indexed.

6.3.3 Random Effects with Spatial Basis Functions

The spatio-temporal process can be written as functions of space basis functions only and their random coefficients are indexed by time. The process is

$$Y(\mathbf{s}; t_j) = \mathbf{x}(\mathbf{s}; t_j)'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha} \phi_i(\mathbf{s})\alpha_i(t_j) + \nu(\mathbf{s}; t_j), \qquad j = 1, \ldots, T. \tag{6.37}$$

where the spatial basis functions are denoted by $\{\phi(\mathbf{s}) : i = 1, \ldots, n_\alpha; \mathbf{s} \in D_{\mathbf{s}}\}$, $\alpha_i(t_j)$ are temporal random processes, and the rest of the model components were defined in 6.34.

In general, one can consider a wide variety of spatial basis functions for this model. These can be of reduced rank, of full rank, or over-complete. One alternative is to consider complete global basis functions such as Fourier. Another option are the reduced-rank empirically defined basis functions, such as the empirical orthogonal functions (EOFs). There are also a variety of non-orthogonal bases for example, Gaussian func-

tions, wavelets, bisquare functions or Wendland functions. It is relevant to ensure that the type and number of basis functions are flexible and large enough to model the true dependence in Y, and the data Z. This requires some experimentation and model diagnostics.

The model 6.37 defined for n locations at times $\{t_j : j = 1, 2, \ldots, T\}$ can be written as

$$\mathbf{Y}_{t_j} = \mathbf{X}_{t_j}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha}_{t_j} + \boldsymbol{\nu}_{t_j}. \tag{6.38}$$

In this case, $\mathbf{Y}_{t_j} = (Y(\mathbf{s}_1; t_j), \ldots, Y(\mathbf{s}_n; t_j))'$ is the n-dimensional process vector,

$$\boldsymbol{\nu}_{t_j} \sim Gau(\mathbf{0}, \mathbf{C}_\nu),$$

$$\boldsymbol{\alpha}_{t_j} \equiv (\alpha_1(t_j), \ldots, \alpha_{n_\alpha}(t_j))',$$

$$\boldsymbol{\Phi} \equiv (\boldsymbol{\phi}(\mathbf{s}_1), \ldots, \boldsymbol{\phi}(\mathbf{s}_n))',$$

and $\boldsymbol{\phi}(\mathbf{s}_i) \equiv (\phi_1(\mathbf{s}_i), \ldots, \phi_{n_\alpha}(\mathbf{s}_i))'$, for $i = 1, \ldots, n$.

If $\boldsymbol{\alpha}_{t_1}, \boldsymbol{\alpha}_{t_2}, \ldots$ are independent in time, where $\boldsymbol{\alpha}_{\boldsymbol{t_j}} \sim iid\ \ Gau(\mathbf{0}, \mathbf{C}_\alpha)$, then the marginal distribution of $\mathbf{Y}_{t_j}$ is $Gau(\mathbf{X}_{t_j}\boldsymbol{\beta}, \boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\nu)$, and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ are independent. The $nT \times nT$ joint spatio-temporal covariance matrix is given by

$$\mathbf{C}_Y = \mathbf{I}_T \otimes (\boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\nu), \tag{6.39}$$

where $\mathbf{I}_T$ is the T-dimensional identity matrix. If independence is assumed in time, it implies a simple separable spatio-temporal dependence structure. To model more complex spatio-temporal dependence structure using exclusively spatial basis functions, one should specify the model for the random coefficients such that $\{\alpha_{tj} : j = 1, \ldots, T\}$ are dependent in time. Under the dynamical approach a conditional temporal dependence can be assumed.

6.3.4 Random Effects with Temporal Basis Functions

Finally, the spatio-temporal random process can be expressed in terms of temporal basis functions and spatially indexed random effects. The process can be written as

$$Y(\mathbf{s}; t) = \mathbf{x}(\mathbf{s}; t)'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha} \phi_i(t)\alpha_i(\mathbf{s}) + \nu(\mathbf{s}; t), \tag{6.40}$$

where the temporal basis functions are denoted by $\{\phi_i(t) : i = 1, \ldots n_\alpha; t \in D_t\}$ and $\{\alpha_i(\mathbf{s}) : \mathbf{s} \in D_\mathbf{s}; i = 1, \ldots, n_\alpha\}$ are their spatially indexed random coefficients. These

coefficients can be modelled using using multivariate geostatistics. The temporal basis function representation is less common than the spatial basis function representation. It could be since most spatio-temporal processes have an interpretation of spatial processes evolving in time. However, temporal basis functions are being used to model non-stationary-in-time processes such as high-frequency time behaviour that vary across space.

There is a spatio-temporal method that implements a low-rank approach to spatial and spatio-temporal modelling known as *fixed rank kriging* (FRK). There are important differences between FRK and other geostatistical implementations. This approach will be followed in the next section applied to LAQN data.

The FKR method can be implemented in **R** through the package FKR, a detailed explanation about the implementation can be seen in Zammit-Mangion & Cressie (2018). The generic basis function that FRK uses by default is the bisquare function, it is given by

$$b(\mathbf{s}, \mathbf{v}) \equiv \begin{cases} \{1 - (||\mathbf{v} - \mathbf{s}||)^2\}^2, & ||\mathbf{v} - \mathbf{s}|| \leq r, \\ 0, & \text{otherwise} \end{cases} \tag{6.41}$$

where r is the aperture parameter. The generic basis function in this spatial covariance model is the local bisquare function, see Cressie & Johannesson (2008). The application for the LAQN data will be performed in the next section.

6.4 Implementation using the LAQN data

When fitting the model 6.34, one of the decisions concerns the choice of basis functions. For spatial processes, this decision is usually not that critical, as there are different types of bases that can accommodate the same spatial variability. However, in the context of spatio-temporal processes, the choice of basis functions can be substantive, especially for the dynamical formulations. In general, one can use

- fixed or parameterised basis functions,
- local or global basis functions,
- reduced-rank, complete, or over-complete bases, and
- basis functions with expansion coefficients possibly indexed by space, time, or space-time.

It can be challenging to come up with good spatio-temporal basis functions. In the same way it has been difficult to come up with realistic spatio-temporal covariance functions.

One option is to consider the tensor product basis functions, they are formally defined in the Appendix A.0.1. The spatio-temporal basis function can be defined as the product of a spatial basis function and a temporal basis function. It is also common to see spatio-temporal-dependence models for Y, where the statistical dependence comes exclusively from spatial basis functions whose coefficients are temporal stochastic processes.

In this section Fixed Rank Kriging (FRK) is applied to the $PM_{2.5}$ data set using space-time tensor product basis functions at two resolutions. In particular, bisquare basis functions are used, these functions were defined by 6.41.

In FRK the basis functions, $\{\phi_i(\mathbf{s};t) : i = 1, \cdots, n_\alpha\}$, are constructed by taking the tensor product of spatial basis functions with temporal basis functions. In particular, let us consider a set of $r_{\mathbf{s}=50}$ spatial functions $\{\phi_p(\mathbf{s}) : p = 1, \ldots, r_{\mathbf{s}}\}$ and a set of $r_t = 60$ temporal basis functions $\{\psi_q(t) : q = 1, \ldots, r_t\}$. Then one construct the set of spatio-temporal basis functions as $\{\phi_{\mathbf{s}t,u}(\mathbf{s},t) : u = 1, \cdots, r_{\mathbf{s}}r_t\} = \{\phi_p(\mathbf{s})\psi_q(t) : p = 1 \cdots r_{\mathbf{s}}; q = 1, \ldots, r_t\}$.

FRK considers the spatio-temporal random-effects model shown in 6.34. It can be applied to the $\log(PM_{2.5})$ data set using $n_\alpha = 3000$ space-time tensor product basis functions at two resolutions for $\{\phi_i(\mathbf{s};t) : i = 1, \cdots, n_\alpha\}$. As in the previous example, the data set to be analysed contains data from 01 August 2015 to 15 March 2016.

The grid cells are known as basic areal units (BAUs), and their primary utility is to account for problems of change of support. Gridded BAUs arranged within a non-convex hull enclosing the data for Greater London area can be seen in figure 6-6. The BAUs are space-time regular lattices, which are classified as a spatio-temporal fixed data frame. In the following example function an R function is used to construct the BAUs in a space-time cube, centred around the data. In particular, each BAU is of size 0.2 deg. latitude $\times$ 0.2 deg. longitude $\times$ 1 day. At this stage, the BAUs only contain geographical information and the number of BAUs depends on the spatial domain boundary. These construction represent some practical advantages such as:

1. Integration of multiple observations with different supports with relative ease.

2. Dimensionality reduction, a relatively small number of basis functions ensures computationally efficient prediction.

3. It is possible to distinguish between measurement error and fine-scale variation at the resolution of the BAU.

It noted that the BAU's are generated automatically using a function contained in the

package FRK. Within the package, it is not possible to generate the same grid that has been used for the rest of the models included in this thesis. This disadvantage is mentioned by the authors as part of future work, see Zammit-Mangion & Cressie (2018).

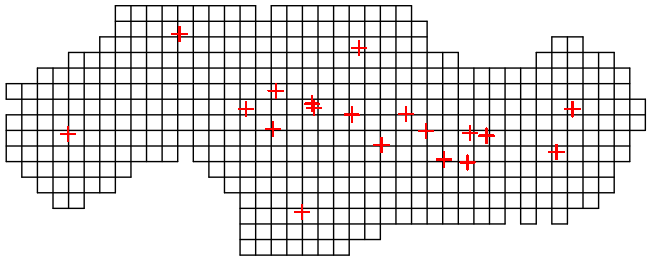

Figure 6-6: BAUs constructed for modelling and predicting $\log(\text{PM}_{2.5})$ levels. The 19 sites in Greater London are highlighted in red.

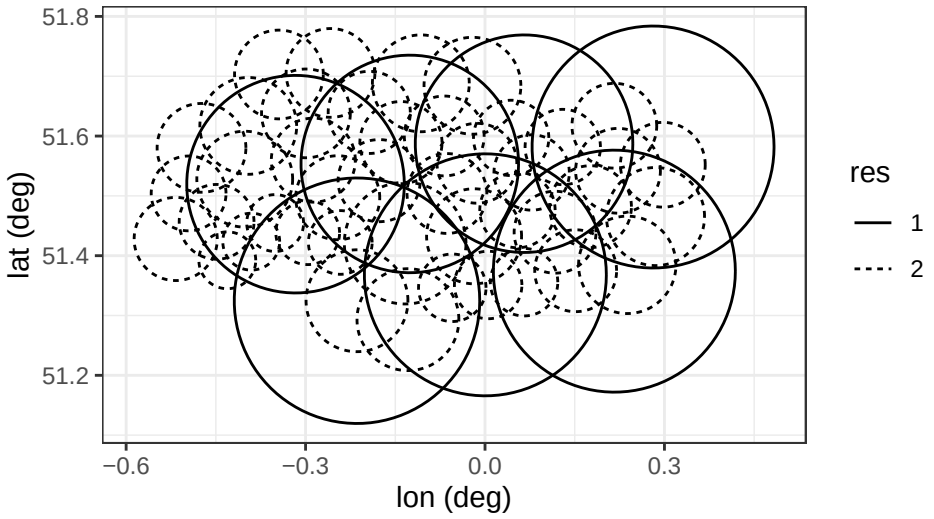

Figure 6-7: Locations of spatial basis functions, circles denote spatial support. The basis functions are automatically generated for $PM_{2.5}$ data set from 01 August 2015 to 15 March 2016 with 2 resolutions. For bisquare functions on the plane, each circle is centred at the basis-function centre, and has a radius equal to the aperture of the function.

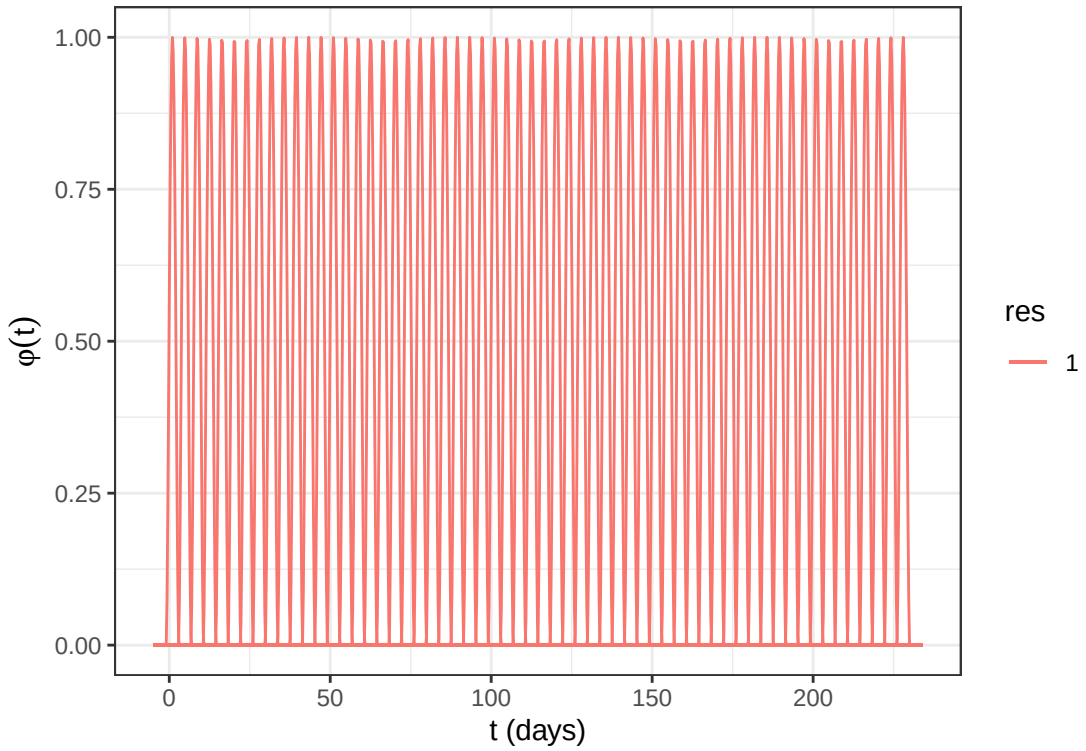

Figure 6-8: Temporal basis functions used to construct the spatio-temporal basis functions. It is a regular sequence of $r_t = 60$ bisquare basis functions between day 1 and day 228 (01 August 2015 - 15 March 2016), the support of each bisquare function is 4 days.

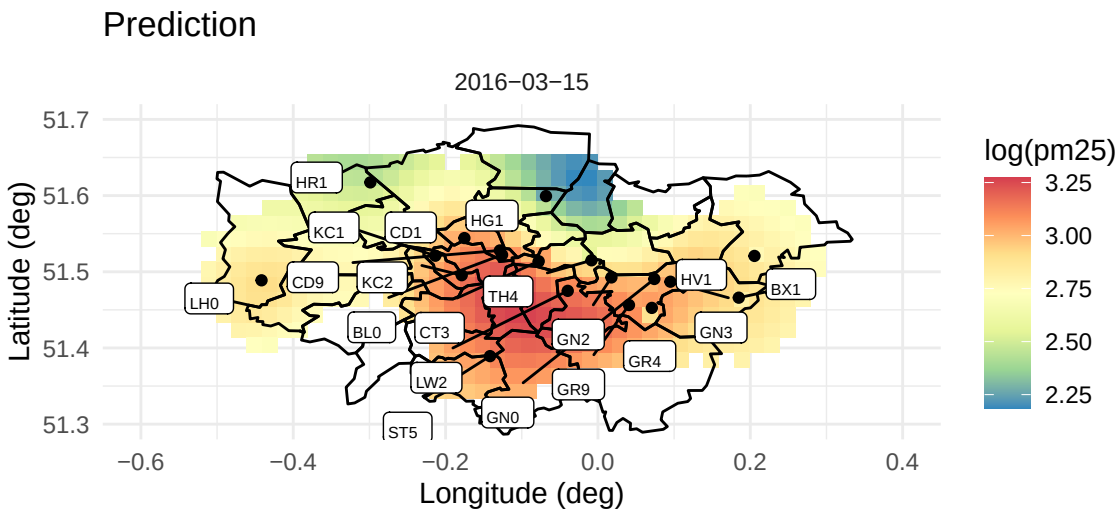

Figure 6-9: FRK predictions of $\log(\text{PM}_{2.5})$ within a square box enclosing the domain of interest for 15 March 2016. Bisquare spatio-temporal basis functions are considered.

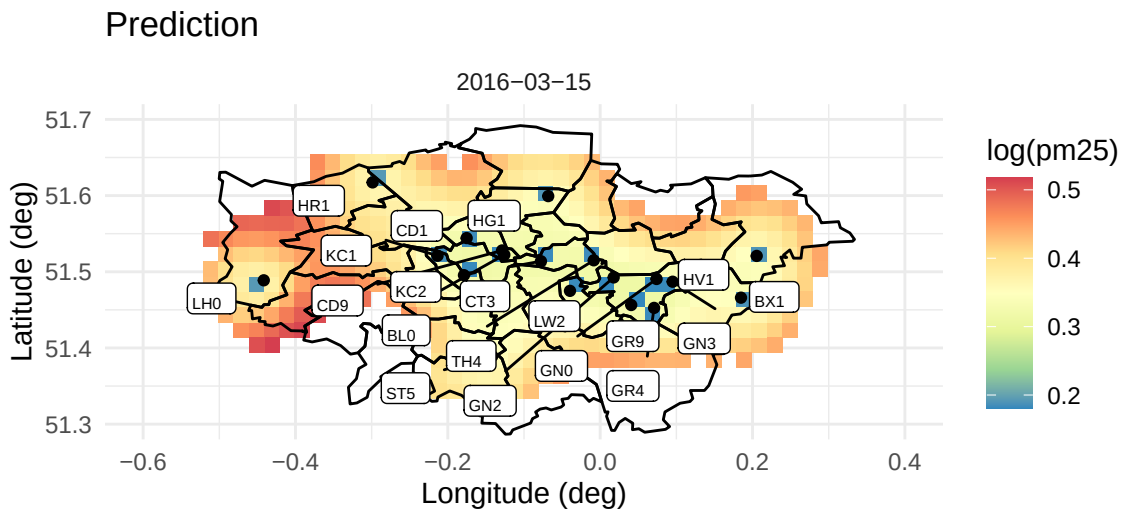

Figure 6-10: FRK prediction standard errors of $\log(\text{PM}_{2.5})$ within a square box enclosing the domain of interest for 15 March 2016. Bisquare spatio-temporal basis functions are considered.

Basis functions can be either regularly placed, or irregularly placed, and they are often multiresolutional. Two resolutions are chosen with $r_{\mathbf{s}} = 50$ spatial basis functions in total, and place them irregularly in the domain. Temporal basis functions also need to be defined. A regular sequence of $r_t = 60$ bisquare basis functions between day 1 and day 227 (01 August 2015 - 15 March 2016) of the month are constructed. In this case, the support of each bisquare function is 4 days. Spatial and temporal basis functions used to construct the spatio-temporal basis functions are shown in Figures 6-7 and 6-8 respectively.

As in the previous universal kriging example, the latitude and longitude are considered as a covariates. The data set analysed is the same, from 01 August 2015 to 15 March 2016. FRK also considers a fine-scale-variation component $\boldsymbol{\nu}$ such that $\mathbf{C}_\nu$ is diagonal. The matrix $\mathbf{C}_\alpha$ is constructed in such a way that the coefficients $\boldsymbol{\alpha}$ at each resolution are independent. The covariances between these coefficients within a resolution decay

exponentially with the distance between the centres of the basis functions. Parameters are estimated using an EM algorithm for computing maximum likelihood estimates. See details in Cressie & Johannesson (2008); Zammit-Mangion & Cressie (2018) and in section 5.9.2.

Figures 6-9 and 6-10 show the predictions and prediction standard errors obtained using FRK, respectively. The predictions are similar to those obtained using Spatio-Temporal kriging in figure 6-4. The prediction standard errors show similar patterns to those observed earlier in figure 6-5, but there are considerable differences based on visual examination. In the context of reduced rank methods, the prediction-standard-error maps can have prediction standard errors related more to the shapes of the basis functions and less to the prediction location's proximity to an observation. The errors associated with those for 15 March are larger nearby to the monitoring sites.

An analogous exercise to the one shown in table 6.1 was made in order to evaluate the spatial predictions obtained from the FRK model that was described above. The idea is fitting the model considering data from 01 August 2015 to 17 March 2016. Instead of using 19 sites from LAQN, three sites are leaved out from the fitting (BL0, GN3, ST5). The same settings used for the 19 sites were considered for this exercise, the only difference is the number of analysed locations. As one can observe in table 6.4, the RMSEs and relative errors are large. This comparison was made in the original scale. It is noticeable that the spatial predictions when fitting this model are over-smoothed. This is not surprising since in this case the prediction is generated into the BAUs. Then, the prediction is the value of the closest grid cell to the monitor locations (BL0, GN3 and ST5). One of the biggest limitations of FRK is that BAUs cannot be easily modified to include specific locations to predict, see Cressie & Johannesson (2008).

6.4.1 Discussion

In this chapter, we have moved from the DSTM approach that was useful in producing temporal predictions/forecasts, but had limitations in making spatial predictions, due to the lack of a defined spatial structure in the model. Here, we describe and implement different methods for Kriging over space and time, that allow both spatial and temporal predictions. The evaluation studies can therefore now include spatial prediction. Spatio-temporal kriging, as shown here, is relatively easy to implement for small data sets. However, it starts to become prohibitive as data sets grow in size, unless some approximation is used. One approach to computational challenges is to use basis function representations and low rank approximations of complex surfaces. In this chapter, Fixed Rank Kriging (FRK) was implemented and compared with the

results from the more traditional approach to Kriging.

In general, a pattern of large errors at the predicted locations was observed for FRK. This is due to predictions being made for the closest grid cell (used in the definition of the BAUs), rather than the exact location of the (omitted from the modelling dataset) locations to be predicted. This is a limitation of the particular implementation used in the package, for which it is not currently possible to easily modify the construction of the BAUs. FRK is designed to work with local basis functions with analytic form. However, this framework could also admit basis functions that have no known functional form. Some examples are empirical orthogonal functions (EOFs) and classes of wavelets defined in an iterative way. Although designed for large data, FRK begins to slow down when a high number of data points are used. In particular, increasing the number of spatio-temporal basis functions could be problematic in terms of computation. In this example, all BAUs are assumed to be of equal area. This is not problematic for this case, since regular grids on the real line and the plane are assumed. However, if there is not an equal area grid, an appropriate weighting should be used when considering arbitrary polygons. The automatic construction of the BAUs could imply some difficulties if one wants to compare them with other models (as in this thesis).

In the next chapter, we move on to using Bayesian hierarchical models that incorporate spatio-dependence and allow predictions over both space and time. In addition to the ability to cover space and time, a major advantage to this approach is that it provides a coherent approach to incorporating uncertainty.

Sites out: BL0, GN3, ST5				
Model	Site	RE: 15/03/2016	RE: 16/03/2016	RE: 17/03/2016
	BL0	17.54%	43.00%	99.47%
FRK	GN3	1.78%	12.75%	99.94%
	ST5	1.04%	11.40%	99.96%
RMSE		11.7696	4.3189	8.9494

Table 6.4: Evaluation (leave-one-out cross validation) of applying the FRK model to data from 01/08/15 to 17/03/16 considering $PM_{2.5}$ data. Results are shown using 16 sites from London data when fitting the model (see text for details). The relative error (RE) per site and date (in percentage) and the overall root mean squared error (RMSE) are shown. Evaluations are presented for 15, 16 and 17 of March 2016. The comparison was made in the original scale ($\mu g/m^3$).

Chapter 7

Bayesian Spatio-Temporal Hierarchical Modelling

Based on the results from chapter 5 it is possible to notice that through this approach it is possible to obtain an accurate forecast for $PM_{2.5}$ levels. It is computationally efficient and by its flexibility, a multivariate model can also be considered to include the simultaneous analysis of more pollutants. However, it should be noticed that the proposed DSTM model does not allow filling in missing data or include uncertainty in prediction. If one is interested in all these purposes, a Bayesian implementation should be considered as an alternative.

Many studies have been developed to the proposal of spatial prediction techniques to obtain concentration maps useful for evaluating human exposure where no measurement stations are located. Some analyses develop methods for temporal forecasting (Sahu *et al.*, 2006) or consider several pollutants at the same time (Shaddick & Wakefield, 2002). Traditional Bayesian methods such as MCMC are computationally limited to inferring spatial and spatio-temporal problems with a large dimension. In recent years, specific finite-dimensional Gaussian random field models have been proposed. A review of recent methods can be found in Cressie & Johannesson (2008); Lindgren *et al.* (2011); Katzfuss (2017).

This chapter outlines a spatio-temporal model for daily air pollution measurements that can be used for short-term forecasting of $PM_{2.5}$. Integrated Nested Laplace Approximation (INLA) will is used in this chapter to fit a latent separable spatio-temporal model. INLA is a Bayesian method that produce approximate marginal (posterior) distributions over all states and parameters. The package **R-INLA** allows for a variety

of modelling approaches (Blangiardo & Cameletti, 2015; Krainski *et al.*, 2018).

The method is illustrated considering air pollution data from Greater London August 2015-March 2016. Different hierarchical spatio-temporal models for particulate matter have been proposed to analyse PM concentration. One of these models consists of a Gaussian Field (GF), affected by a measurement error, and a state process. This state process can be characterised by a first order autoregressive dynamic model and spatially correlated innovations, see Cameletti *et al.* (2013); Blangiardo & Cameletti (2015). In these studies a Stochastic Partial Differential Equation (SPDE) approach is used for a hierarchical spatio-temporal models for particular matter concentration in the North-Italian region Piemonte. The same methodology will be followed in this thesis.

In general, a GF (Rue & Held, 2005) is completely specified by its mean and spatio-temporal covariance function. Even if a GF is easily defined through its first and second moments, the implementation can be limited due to its high dimensionality. If our objective is dealing with large data sets in space and time, there is a big computational cost required for model fitting and spatial interpolation and prediction.

The idea behind this approach is represent GF as a Gaussian Markov Random Field (GMRF) (Rue & Held, 2005; Chung, 2020). A GMRF is a spatial process that models the spatial dependence of data observed on areal units, such as regular grid or lattice structure. For further details, see Rue & Held (2005); Cameletti *et al.* (2013). The objective of this approach is to find a GMRF. It can be made trough a local neighbourhood and sparse precision matrix, that best describes the Matérn field. This representation has some computational properties that can be used to make inference in the GMRF. So that, the high dimensionality problem that arises when working with the dense covariance matrix of a GF can be avoided.

7.1 A spatio-temporal model for air pollution

As described in chapter 1, some examples of the use of INLA can be found in the context of air pollution analysis. Cameletti *et al.* (2011) employed the SPDE approach for a hierarchical spatio-temporal models for particular matter concentration in the North-Italian region Piemonte in winter season. The model involves a Gaussian field, affected by a measurement error, and a state process characterised by a AR(1) dynamics and spatially correlated innovations. The parameter posterior estimates are obtained together with prediction and uncertainty map. Moreover, the posterior distributions over the triangulated domain are also generated. They conclude that it is an acceptable

computational cost of the SPDE approach implemented by INLA algorithm compared with other approaches.

Let us consider

$$\{Y(\mathbf{s},t) : \mathbf{s} \in D_{\mathbf{s}}, t \in D_t\} \tag{7.1}$$

as the spatio-temporal process (or random field) that is a statistical model for a phenomenon. It can be assumed $D_{\mathbf{s}} \subseteq \mathbb{R}^2$ and $D_t \subseteq \mathbb{R}$; for convenience, the notation can differ slightly from previous sections. These realisations are used to make inference about the process and to predict it at desired locations.

Let $Y(\mathbf{s}_i,t)$ denote the realisation of the spatio-temporal process $Y(\cdot,\cdot)$ previously defined. It represents the $PM_{2.5}$ concentration at station $i = 1,\ldots,m$ located at location $\mathbf{s}_i$ and day $t = 1,\ldots,T$. It can be written as

$$Y(\mathbf{s}_i,t_j) = \mathbf{x}(\mathbf{s}_i,t_j)\boldsymbol{\beta} + \xi(\mathbf{s}_i,t_j) + \epsilon(\mathbf{s}_i,t_j), \tag{7.2}$$

where $\mathbf{x}(\mathbf{s}_i,t_j) = (x_1(\mathbf{s}_i,t_j),\ldots,x_p(\mathbf{s}_i,t_j))$ denotes the vector of p covariates at location $\mathbf{s}_i$ at time t_j, and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ is the coefficient vector. This model could include covariates related to air pollution such as temperature, precipitation or distance to roads, and this will improve predictive performance. The realisation of the state process is denoted by $\xi(\mathbf{s}_i,t_j)$. In this case, $\{\epsilon(\mathbf{s}_i,t_j)\}$ is independent of $Y(\cdot;\cdot)$ and represent the measurement error that is assumed to be *iid* with mean zero and variance σ^2_ϵ.

If one is analysing air pollution data, the true unobserved level of pollution is assumed to be a spatio-temporal process that changes in time with first order autoregressive dynamics. It is given by

$$\xi(\mathbf{s}_i,t_j) = a\xi(\mathbf{s}_i,t_{j-1}) + \omega(\mathbf{s}_i,t_j) \tag{7.3}$$

for $j = 2,\ldots,T$, where $|a| < 1$ and $\xi(\mathbf{s}_i,t_1)$ is obtained from the stationary distribution $N(0,\sigma^2_\omega/(1-a^2))$. In this case, $\omega(\mathbf{s}_i,t_j)$ has a zero-mean Gaussian distribution and is assumed to be temporally independent. It is characterised by the spatio-temporal covariance function,

$$\text{cov}(\omega(\mathbf{s}_i,t_j),\omega(\mathbf{s}_k,t_l)) = \begin{cases} 0 & \text{if } j \neq l, \\ C(d) & \text{if } j = l \end{cases} \tag{7.4}$$

defined for $i \neq k$. The function $C(d)$ is the Matérn spatial covariance function defined in 4.14. It depends on the location $\mathbf{s}_i$ and $\mathbf{s}_k$ only through the Euclidean spatial distance

$d = ||\mathbf{s}_i - \mathbf{s}_k|| \in \mathbb{R}$. Thus, the process is second-order stationary and isotropic. The spatio-temporal covariance function in equation 7.4 is separable, i.e. it can be rewritten as the product of a purely spatial and a purely temporal covariance function.

Let us denote all the observations measured at time t by

$$\mathbf{Y}_t \equiv (Y(\mathbf{s}_1, t), \ldots, Y(\mathbf{s}_m, t))'. \tag{7.5}$$

It follows that 7.2 and 7.3 can be rewritten as

$$\mathbf{Y}_t = \mathbf{x}_t\boldsymbol{\beta} + \boldsymbol{\xi}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_m) \tag{7.6}$$

$$\boldsymbol{\xi}_t = a\boldsymbol{\xi}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma} = \sigma_\omega^2 \tilde{\boldsymbol{\Sigma}}), \tag{7.7}$$

where $\mathbf{I}_m$ is the identity matrix of dimension m, $\mathbf{x}_t = (x(\mathbf{s}_1, t), \ldots, x(\mathbf{s}_m, t))$ and $\boldsymbol{\xi}_t = (\xi(\mathbf{s}_1, t), \ldots, \xi(\mathbf{s}_m, t))$. In particular, $\boldsymbol{\xi}_1$ corresponds to the stationary distribution of the AR(1) process, $N(\mathbf{0}, \boldsymbol{\Sigma}/(1-a^2))$. The correlation matrix has dimension m and is denoted by $\tilde{\boldsymbol{\Sigma}}$ with elements $C(||\mathbf{s}_i - \mathbf{s}_j||)$, where $C(\cdot)$ is the Matérn function given by 4.14 and is parameterised by κ and ν.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\epsilon^2, a, \sigma_\omega^2, \kappa)$ denote the parameter to be estimated. The joint posterior distribution is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\xi}, \boldsymbol{\theta})\pi(\boldsymbol{\xi}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{7.8}$$

where $\mathbf{Y} = \{\mathbf{Y}_t\}$ and $\boldsymbol{\xi} = \{\boldsymbol{\xi}_t\}$ with $t = 1, \ldots, T$.

In a Bayesian statistics, it is common practice making inference from 7.8 by MCMC sampling, see Cameletti *et al.* (2013) for implementation details. One alternative to MCMC is representing a GF with Matérn covariance function as a GRMF. For theoretical details and proofs of this alternative see Lindgren *et al.* (2011).

7.2 Gaussian Markov Random Fields GMRFs

A GMRF is a spatial process that models the spatial dependence of data observed on areal units, regular grid or lattice structure, see Rue & Held (2005). Let us denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ as an n-dimensional GMRF with mean μ and symmetric and positive definite precision matrix $\mathbf{Q}$ (the inverse of the covariance

matrix). The density is given by

$$\pi(\mathbf{Y}) = (2\pi)^{-n/2}|\mathbf{Q}|^{1/2}\exp\left(-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{Q}(\mathbf{Y}-\boldsymbol{\mu})\right). \tag{7.9}$$

A GMRF $\mathbf{Y}$ can be specified through the conditional distributions for each component given all the others. And due to the Markovian property, the full conditional distribution of Y_i $(i = 1, \ldots, n)$ depends only on a few of the components of $\mathbf{Y}$. This set of components is denoted by δ_i, which is composed by a set of neighbours of unit i and

$$\pi(\mathbf{Y}_i|\mathbf{Y}_{-i}) = \pi(Y_i|\mathbf{Y}_{\delta_i}), \tag{7.10}$$

where $\mathbf{Y}_{-i}$ denotes all elements in $\mathbf{Y}$ except for Y_i. Here, if one follows the notation of Rue and Held (2005), one can say that that given the neighbourhood δ_i, the terms Y_i and $Y_{-\{i,\delta_i\}}$ are independent. This conditional independence relation can be written as

$$Y_i \perp \mathbf{Y}_{-\{i,\delta_i\}}|Y_{\delta_i} \tag{7.11}$$

for $i = 1, \ldots, n$. It is important that this conditional independence property is strictly determined by the precision matrix $\mathbf{Q}$. This matrix will be sparse due to a conditional independence property and one can take advantage in terms of computation. Let us consider a general couple i and j with $i \neq j$, it holds that

$$Y_i \perp Y_j|\mathbf{Y}_{-\{i,j\}} \iff \mathbf{Q}_{ij} = 0. \tag{7.12}$$

It means that the nonzero pattern of $\mathbf{Q}$ is given by the neighbourhood structure of the process. That is, $\mathbf{Q}_{ij} \neq 0$ if $j \in \{i, \delta_i\}$.

7.3 The stochastic partial differential equations (SPDEs) approach

Let

$$Y(\mathbf{s}) \equiv \left\{Y(\mathbf{s}) : \mathbf{s} \in D_{\mathbf{s}} \subseteq \mathbb{R}^2\right\} \tag{7.13}$$

a second-order stationary and isotropic Gaussian field (GF), which spatial correlation function is defined by a Matérn function. In particular, the correlation function is defined by

$$C(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^{\nu}K_{\nu}(\kappa h) \tag{7.14}$$

where K_ν denotes the modified Bessel function of the second kind and order $\nu > 0$. The parameter ν is usually kept fixed and $\nu > 0$ is a smoothness parameter.

Let us suppose we have a realisation of the process $\mathbf{Y}(\mathbf{s}_i)$ at m spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$. The idea behind of the SPDE approach is to find a GMRF, with local neighbourhood and sparse precision matrix $\mathbf{Q}$. The objective in this context of this model is trying to find this matrix that best represents the Mátern field. Under these assumptions it is possible to make inference using the GMRF taking advantage of its computational properties. Thus, it is possible to avoid the high dimensionality problem that arises when working with the dense covariance matrix of a GF.

The SPDE approach considers a finite element representation to define the Matérn field as a linear combination of basis functions defined on a triangulation of the domain $D_{\mathbf{s}}$. The triangulation consists in subdividing $D_{\mathbf{s}}$ into a set of non-intersecting triangles meeting in at most a common edge. At the beginning, the triangle initial vertices are placed at the locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$. Then, new vertices can be added to get a triangulation useful for spatial prediction purposes. To illustrate the concept of triangulation, an example is provided referring to the London AQN data previously discussed. The figure 7-1 displays the locations of the 19 $PM_{2.5}$ monitoring stations (in blue) with their respective triangulation of the region using 105 vertices. Once a triangulation of the domain has been defined, the spatial field can be approximated by

$$\mathbf{Y}(\mathbf{s}) = \sum_{l=1}^{n} \Psi_l(\mathbf{s})\omega_l, \tag{7.15}$$

where n is the total number of vertices (or nodes), $\{\Psi_l(\mathbf{s})\}$ are the basis function and $\{\omega_l\}$ are Gaussian distributed weights. To ensure the Markov structure that is required for a GMRF, the set of basis functions should be piecewise linear,

$$\Psi_l(\mathbf{s}) = \begin{cases} 1 & \text{at vertex } l\ , \\ 0 & \text{at all other vertices.} \end{cases} \tag{7.16}$$

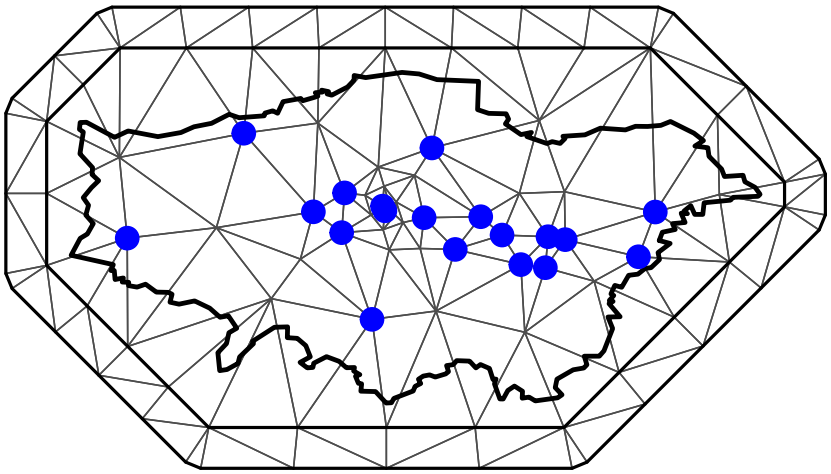

Figure 7-1: Triangulation for the locations of monitoring sites within the Greater London area for use with the SPDE approach to modelling spatio-temporal data with INLA. The mesh comprises 105 vertices and the monitoring locations are highlighted in blue.

Through the piecewise linear basis functions, one can represent a continuous spatial random field in its corresponding finite element representation. The value of the spatial field at each triangle vertex is given by the weight ω_l. And the values in the interior of the triangle are determined by linear interpolation.

The idea behind the SPDE approach is establishing the link between the GF, $\mathbf{Y}(\mathbf{s})$, and the GMRF. This can be done with the Gaussian weights, ω_l in 7.15 and assuming the Markovian structure, see Lindgren *et al.* (2011).

Let us consider a Matérn GF, $\mathbf{Y}(\mathbf{s})$, defined by 7.14, then the representation given by 7.15 is a finite element method solution of a SPDE. The solution is given by

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau\omega_{\mathbf{s}}) = \mathcal{W}_s \quad \mathbf{s} \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0, \tag{7.17}$$

where $(\kappa^2 - \Delta^{\alpha/2})$ is a pseudodifferential operator, Δ is the Laplacian, κ is the scale parameter, τ controls the variance, ν the smoothness and $\mathcal{W}_{\mathbf{s}}$ is spatial white noise with unit variance. Any GRF model defined with a Matérn covariance structure can be approximated by a GMRF, $\nu+d/2$ is integer valued. This method can be extended to GRFs on manifolds, non-stationary and anisotropic covariance structures. For further details see Lindgren *et al.* (2011); Krainski *et al.* (2018).

Now, the interest is on the resulting precision matrix of the observations $\mathbf{Q}_{\alpha,\kappa}$. Where the triangulation and the basis functions are considered. Let us suppose the $m \times m$ matrices $\mathbf{C}$, $\mathbf{G}$ and $\mathbf{K}_\kappa$. Their respective entries are

$$\begin{aligned} \mathbf{C}_{ij} &= \langle \psi_i, \psi_j \rangle, \\ \mathbf{G}_{ij} &= \langle \Delta\psi_i, \Delta\psi_j \rangle, \\ (\mathbf{K}_\kappa)_{i,j} &= \kappa^2 C_{i,j} + G_{i,j}. \end{aligned}$$

Where $\langle \cdot, \cdot \rangle$ denotes the inner product and Δ the gradient. The precision matrix $\mathbf{Q}_{\alpha,\kappa}$ as a function of κ^2 and α can be expressed as

$$\mathbf{Q}_{1,\kappa} = \mathbf{K}_\kappa = \kappa^2\mathbf{C} + \mathbf{G}, \tag{7.18}$$

$$\mathbf{Q}_{2,\kappa} = \mathbf{K}_\kappa\mathbf{C}^{-1}\mathbf{K}_\kappa = \kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}, \tag{7.19}$$

$$\mathbf{Q}_{\alpha,\kappa} = \mathbf{K}_\kappa\mathbf{C}^{-1}Q_{\alpha-2,\kappa}\mathbf{C}^{-1}\mathbf{K}_\kappa, \quad \text{for} \quad \alpha = 3, 4, \ldots \tag{7.20}$$

Due to the matrix $\mathbf{C}$ is dense, it can be replaced by a diagonal matrix $\tilde{\mathbf{C}}$ with

$$\tilde{\mathbf{C}} = \langle \psi_i, 1 \rangle.$$

In particular, since $\tilde{\mathbf{C}}$ is diagonal, $\mathbf{K}_\kappa$ is sparse as $\mathbf{G}$.

7.4 Integrated Nested Laplace Approximations

For each time point $t = 1, \ldots, T$, the Matérn field, $\boldsymbol{w}_t$, shown in 7.7 is represented through the GMRF representation. Explicitly it is defined by

$$\tilde{\boldsymbol{\omega}}_t \sim N(\mathbf{0}, \mathbf{Q}_\mathbf{s}^{-1}), \tag{7.21}$$

where the precision matrix, $\mathbf{Q_s}$ is obtained from the SPDE representation previously discussed, see 7.18. A detailed explanation of full computation of this matrix is given in Lindgren *et al.* (2011). The matrix $\mathbf{Q_s}$ is time independent and its dimension n is given by the number of vertices of the domain triangulation. The equation 7.7 can be written as

$$\boldsymbol{\xi}_t = a\boldsymbol{\xi}_{t-1} + \tilde{\boldsymbol{\omega}}_t, \qquad \tilde{\boldsymbol{\omega}}_t \sim N(\mathbf{0}, \mathbf{Q}_\mathbf{s}^{-1}). \tag{7.22}$$

for $t = 1, \ldots, T$ and $\boldsymbol{\xi}_1 \sim N(\mathbf{0}, \mathbf{Q}_\mathbf{s}^2/(1-a^2))$. Then, the joint distribution of the Tn-dimensional GMRF $\boldsymbol{\xi} = (\boldsymbol{\xi}_t^{'}, \ldots, \boldsymbol{\xi}_T^{'})$ is

$$\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{Q}^{-1}), \tag{7.23}$$

with $\mathbf{Q} = \mathbf{Q}_T \otimes \mathbf{Q_s}$. Where $\mathbf{Q}_T$ is the T-dimensional precision matrix of the temporal autoregressive process of order 1 specified by 7.22. The matrix is given by

$$\mathbf{Q}_T = \begin{pmatrix} 1/\sigma_\omega^2 & -a/\sigma_\omega^2 & & & \\ -a/\sigma_\omega^2 & (1+a^2)/\sigma_\omega^2 & & & \\ & & \ddots & & \\ & & & (1+a^2)/\sigma_\omega^2 & -a/\sigma_\omega^2 \\ & & & -a/\sigma_\omega^2 & 1/\sigma_\omega^2 \end{pmatrix} \tag{7.24}$$

The equation 7.6 can be rewritten as

$$\mathbf{Y}_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\xi}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_d), \tag{7.25}$$

where $\mathbf{B}$ is a matrix with dimension $m \times n$. This matrix specifies the value of the GRMF, $\boldsymbol{\xi_t}$, for each vector $\mathbf{Y}_t$. The matrix $\mathbf{B}$ is sparse and with only one unit element

for each row and such that

$$Y(\mathbf{s}_i, t) = \mathbf{x}(\mathbf{s}_i, t)\boldsymbol{\beta} + \sum_{j=1}^{n} \mathbf{B}_{ij}\boldsymbol{\xi}_t + \epsilon(\mathbf{s}_i, t), \tag{7.26}$$

where $\mathbf{B}_{ij} = 1$ if the triangle vertex j is placed at location $\mathbf{s}_i$ and 0 elsewhere.

The equations 7.25 and 7.22 define a hierarchical model that is a subclass of structured additive regression models, known as latent Gaussian models. These type of models can be estimated using the integrated nested Laplace approximations (INLA) algorithm proposed in Rue *et al.* (2009). INLA is a computational approach for Bayesian inference and is an alternative to MCMC. It consists in getting the approximated posterior marginals for the latent variables as well as for the hyperparameters.

Let $\boldsymbol{\Psi} = \{\boldsymbol{\beta}, \boldsymbol{\xi}\}$ be the underlying latent field with a priori independent components. A vague Gaussian prior with known precision is assigned to $\boldsymbol{\beta}$ and the GMRF distribution defined in 7.23 to $\boldsymbol{\xi}$. Then the density $\pi(\boldsymbol{\Psi}|(\sigma^2_\omega, a, \kappa))$ is Gaussian with zero mean and precision matrix $\mathbf{Q}$. Notice that the observations $\mathbf{Y} = \{\mathbf{Y}_t\}$ are normally distributed and conditionally independent given $\boldsymbol{\Psi}$ and σ^2_ϵ. If the hyperparameter is denoted by $\boldsymbol{\theta} = (\sigma^2_\omega, a, \kappa, \sigma^2_\epsilon)$, the joint posterior distribution is given by

$$\pi(\boldsymbol{\Psi}, \boldsymbol{\theta}|\mathbf{Y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\Psi}|\boldsymbol{\theta}) \prod_{t=1}^{T} \pi(\mathbf{Y}_t|\boldsymbol{\Psi}, \boldsymbol{\theta}), \tag{7.27}$$

where $\pi(\mathbf{Y}_t|\boldsymbol{\Psi}, \boldsymbol{\theta}) \sim N(\mathbf{x}_t\beta + \mathbf{B}\boldsymbol{\xi}_t, \sigma^2_\epsilon \mathbf{I}_d)$ is the conditional distribution of the $PM_{2.5}$ observations at time t defined by 7.25.

One can be interested in producing a set of predictions of air pollution over an area and in time. With an abuse of notation, let $\hat{Y}_{\mathbf{s}t}$ denote a prediction in a particular location, $\mathbf{s}$, and time, t. The marginal posterior distribution for a prediction can be written as

$$\pi(\hat{Y}_{st}|\mathbf{Y}) = \int\int \pi(\hat{Y}_{st}|\boldsymbol{\Psi}, \boldsymbol{\theta}, \mathbf{Y})\pi(\boldsymbol{\Psi}|\boldsymbol{\theta}, \mathbf{Y})\pi(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\Psi}d\boldsymbol{\theta}. \tag{7.28}$$

One could also be interested in finding the marginal posterior densities for each θ_i and ψ_j given the observed data $\mathbf{Y}$,

$$\begin{aligned} \pi(\psi_i|\mathbf{Y}) &= \int \pi(\psi_i|\boldsymbol{\theta}, \mathbf{Y})\pi(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta} \\ \pi(\theta_j|\mathbf{Y}) &= \int \pi(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}_{-j} \end{aligned} \tag{7.29}$$

for $i = 1, \ldots, T + p$ and $j = 1, \ldots, 4$.

In this case, $\boldsymbol{\theta}_{-i}$ denotes the set of parameters, $\boldsymbol{\theta}$, with the i-th entry removed. In most of the cases, these distributions are not be analytically tractable. Although the INLA algorithm originally was designed for non-Gaussian responses, it can substitute MCMC simulations. The approximations to the distributions in 7.29 are denoted by $\hat{\pi}(\psi_i|\mathbf{Y})$ and $\hat{\pi}(\theta_j|\mathbf{Y})$ respectively. For this particular case, one has that $\hat{\pi}(\psi_i|\mathbf{Y})$ is exact and Gaussian. For computing $\hat{\pi}(\theta_j|\mathbf{Y})$, a numerical integration approximation is required.

The approximation for $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$, is given by

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) \propto \left.\frac{\pi(\mathbf{Y}, \boldsymbol{\Psi}, \boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{\Psi}|\boldsymbol{\theta}, \mathbf{Y})}\right|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}(\boldsymbol{\theta})},$$

where $\tilde{\pi}(\boldsymbol{\Psi}|\boldsymbol{\theta}, \mathbf{Y})$ is a Gaussian approximation of $\pi(\boldsymbol{\Psi}|\boldsymbol{\theta}, \mathbf{Y})$ evaluated at the mode $\hat{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ of the distribution $\boldsymbol{\Psi}|\boldsymbol{\theta}$. The approximation, $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$, is equivalent to a Laplace approximation, and it is exact if $\tilde{\pi}(\boldsymbol{\Psi}|\boldsymbol{\theta}, \mathbf{Y})$ is Gaussian. The approximation used for the posterior, $\pi(\psi_j|\boldsymbol{\theta}, \mathbf{Y})$ is given by

$$\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \mathbf{Y}) \propto \left.\frac{\pi(\mathbf{Y}, \boldsymbol{\Psi}, \boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{\psi}_{-j}|\psi_j, \boldsymbol{\theta}, \mathbf{Y})}\right|_{\boldsymbol{\Psi}_{-j}=\hat{\boldsymbol{\Psi}}_{-j}(\psi_j,\boldsymbol{\theta})},$$

where $\tilde{\pi}(\boldsymbol{\Psi}_{-j}|\psi_j, \boldsymbol{\theta}, \mathbf{Y})$ is a Gaussian approximation of the distribution $\pi(\boldsymbol{\Psi}_{-j}|\psi_j, \boldsymbol{\theta}, \mathbf{Y})$. The distribution $\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \mathbf{Y})$ is obtained by taking Taylor expansions of $\pi(\mathbf{Y}, \boldsymbol{\psi}, \boldsymbol{\theta})$ and $\tilde{\pi}(\boldsymbol{\Psi}_{-j}|\psi_j, \boldsymbol{\theta}, \mathbf{Y})$, up to third order, aiming to correct a Gaussian approximation for location errors due to potential skewness (Rue *et al.*, 2009).

To estimate the marginal posterior distributions given by in 7.29, a set of integration points and weights are built using the distribution $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$. Initially, one have to find numerically by Newton-type algorithms the mode of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$. Around the mode, the distribution $\log(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}))$ is evaluated over a grid of K points $\{\boldsymbol{\theta}^{(k)}\}$, each with associated integration weights $\{\Delta^{(k)}\}$. If the points define a regular lattice, then the integration weights will be equal. The marginal posteriors, $\pi(\theta_k|\mathbf{Y})$, are obtained using numerical integration of an interpolant of $\log(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}))$. The marginal posteriors, $\pi(\psi_j|\mathbf{Y})$, are obtained using numerical integration

$$\pi(\psi_j|\mathbf{Y}) = \sum_k \tilde{\pi}(\psi_j|\boldsymbol{\theta}^{(k)}, \mathbf{Y})\tilde{\pi}(\boldsymbol{\theta}^{(k)}|\mathbf{Y})\Delta^{(k)}, \tag{7.30}$$

where $\tilde{\pi}(\psi_j|\boldsymbol{\theta}^{(k)}, \mathbf{Y})$ and $\tilde{\pi}(\boldsymbol{\theta}^{(k)}|\mathbf{Y})$ are the posterior distributions $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$ and $\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \mathbf{Y})$ evaluated at the set of integration points $\{\boldsymbol{\theta}^{(k)}\}$ while $\Delta^{(k)}$ are integration weights, see (Martins *et al.*, 2013).

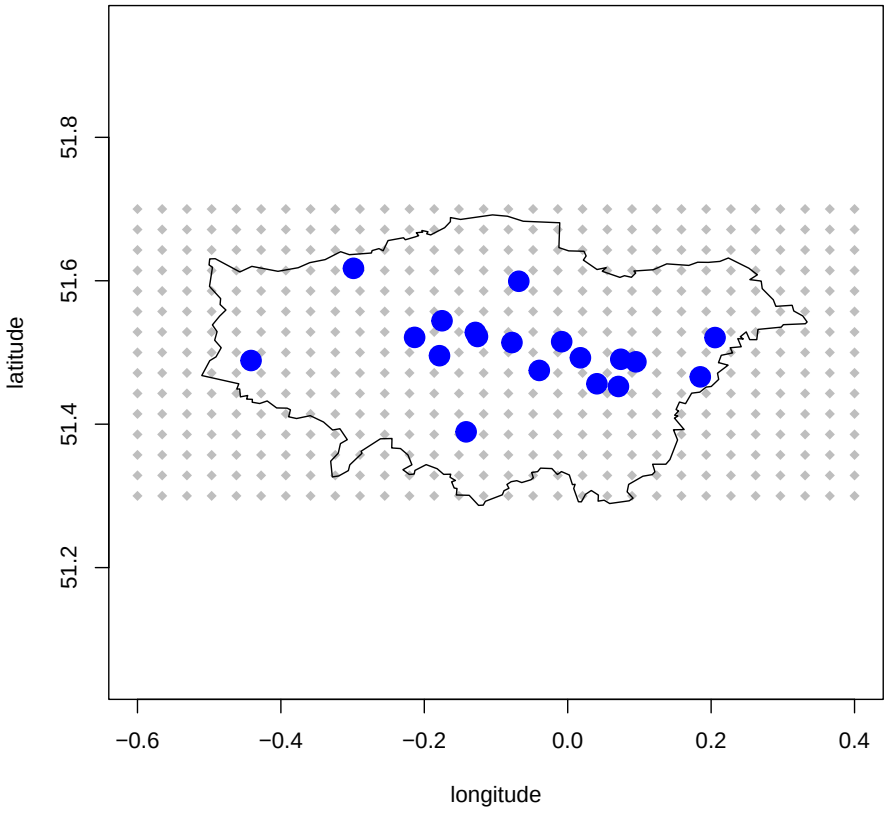

Figure 7-2: Grid for Greater London area, the blue dots denote the 19 monitoring stations considered for a SPDE–INLA model.

7.5 Implementation using the LAQN data

The model presented in section 7.1 is now implemnented using the London AQN data. It is coded in the **R** package INLA, and based on the integrated nested Laplace approximation (INLA) method. It has proven to be a valid alternative to the commonly used Markov Chain Monte Carlo (MCMC) simulations.

7.5.1 Evaluation strategy

In order to assess the accuracy of the BHSTM approach to producing predictions of levels of $PM_{2.5}$ over time and space, we again perform a series of studies and evaluate how well they perform, based upon the RMSE. Temporally, the basis of the evaluations is similar to that presented in 5.10.1 and 6.2.1. In the first instance, we use data on

$PM_{2.5}$ levels from 01 August 2015 to 14 March 2016 and predict for 15, 16 and 17 March 2016. This is repeated for the other time periods described in 5.10.1. The sensitivity of the results to the parameters in the model (section 7.1) is assessed by running the model with a selection of different values for α, the smoothness parameter and the sensitivity to the choice of priors for κ, the scale parameter. Furthermore, τ, the variance parameter (equations 7.17 and 7.14) is also assessed. In assessing the ability to forecast temporally at unmeasured locations, again three spatial locations were omitted from the data sets used to fit the model (sites BL0, GN3 and ST5) and then (out-of-sample) predictions three-days-ahead in both space and time are calculated for the three locations.

7.6 Results

Three different analyses have been performed for this section assuming the same model with different settings and assumptions. In the analyses, the model is fitted using data from different dates up to 14 March 2016. To run the studies for the model with INLA, $\log(PM_{2.5})$ data from sites in London Greater London were considered. This model includes an intercept and a spatio-temporal random effect that changes in time with first order autoregressive dynamics and spatially correlated innovations. However it does not include covariates or information from other pollutants. The same triangulation of the region using 105 vertices was used to run the first two exercises, see figure 7-1. It should be noticed that increasing the number of vertices implies higher computational cost with no benefits in terms of performance. The first step to fit an SPDE model is theconstruction of the mesh. This step must be done carefully. When setting the mesh, the points and a set of constraints need to be passed. The shape of the boundary can be controlled, including its convexity, concavity and resolution. For LAQN data, different mesh configurations were run with similar results in terms of prediction. In this instance, the results were not found to be sensitive to the choice of mesh. These three studies are explained below.

The first study is similar to the ones described in chapter 5 and the results are comparable. To run this study, $\log(PM_{2.5})$ data from 19 site coordinates in London Greater London were considered. The idea is forecasting three days ahead at the 19 locations. This study is proposed in order to evaluate the accuracy of the INLA model in terms of prediction. The forecast for 15, 16, and 17 March will be produced under the INLA approach using different days of data as input. The study consists in analysing the results for three different periods of time. The first interval of time corresponds to the period from 01 August 2015 to 14 March 2016 (227 days). The sensitivity of the results

to the parameters in the implementation of the spatio-temporal model (section 7.1) is assessed by running the model with a selection of different values for α, the smoothness parameter and the sensitivity to the choice of priors for κ, the scale parameter, and τ, the variance parameter (equations 7.17 and 7.14) is also assessed.

As before, some conventions will be followed for the column names in the tables with results, see table 7.1. The site column denotes the code from LAQN (see table 2.1) and the real column represents the actual value at each site. Every column with the the value for $\alpha = 2, 1.5$ will represent the results for these INLA settings, see 7.17. The multicolumn 7 months refers to the period from 01 August 2015 to 14 March 2016. On the other hand, the period from 01 December 2015 to 14 March 2015 is represented by the 3 months multicolumn. And, the 2 months multicolumn corresponds to the period from 01 January 2016 to 14 March 2016. Finally, the last row for each table shows the RMSE for each predicted time. The result for this study can be seen in the table 7.1.

Some general observations can be made from the table 7.1. In terms of RMSE, there is not a large difference when using 7 and 3 months of data. The smallest errors are observed for 2 months of observation. In the same way, a small difference is observed between the forecast for $\alpha = 2$ and $\alpha = 1.5$. But, in general smaller errors are observed for $\alpha = 1.5$. The best prediction for 15 March 2016 is when $\alpha = 1.5$ for 2 months, the error is similar when $\alpha = 2$ for the same period of time. The best prediction for 16 March 2016 is observed when fitting the model with $\alpha = 1.5$ for 2 months. On the other hand, the best result for 17 March 2016 is when $\alpha = 1.5$ for 2 months. There is an important variation in accuracy for the prediction for different sites. For instance, the biggest difference in error percentage (not shown in the tables) for 15 March 2016 is found in the site HG1. As one would expect in this type of model, the RMSE rises when more days ahead are predicted.

In order to compare the previous results with two more time intervals. In particular, the result forecast is done for 26, 27 and 28 March 2016. The three periods of time starts on 12 August 2015 (227 days), 12 December 2015 (105 days) and 01 January 2015 (74 days). In particular, all the models are fitted using up to 25 March 2016 data. The results considering the values $\alpha = 2, 1.5$ are found in the table 7.2. In terms of RMSE, the difference between using 7, 3 and 2 months of data is relatively small. In general, a small difference is observed when $\alpha = 2$ or $\alpha = 1.5$. In this case, the best prediction for the three days ahead is reached when fitting $\alpha = 1.5$ for 2 months.

An analogous study is also proposed. In this case the result forecast is done for 14, 15 and 16 October 2015. The three periods of time starts on 01 March 2015 (227

days), 01 July 2015 (105 days) and 01 August 2015 (74 days). In particular, all the models are fitted using up to 13 October 2015 data. The results considering the values $\alpha = 2, 1.5$ are found in the table 7.3. In terms of RMSE, the difference between using 7, 3 and 2 months of data is slightly smaller than the previous results. In general, a small difference is observed when $\alpha = 2$ or $\alpha = 1.5$. The best prediction for 14 October 2015 is when $\alpha = 1.5$ for 2 months. The best prediction for 15 October 2015 is observed when fitting $\alpha = 1.5$ for 7 months. Finally, the best result for the 16 October 2015 is when $\alpha = 1.5$ for 3 months. In contrast to the previous table, the RMSEs are larger for one day ahead than for two days ahead forecast. There is also an important variation in accuracy for the prediction for different sites. In general, RMSEs are larger for this study than those ones from the previous analyses.

In the second study the interest is in the prediction of $PM_{2.5}$ concentration on a regular grid for 15, 16 and 17 March 2016. This prediction was made over a 30×15 regular grid for London Greater area, it is shown in figure 7-2. A posterior distribution is generated in contrast with the previous approaches that generate a point estimation (DSTM). Depending on the final purpose of the model, having a posterior distribution and a probability interval can be more convenient. The forecast using 7 months of data for 15 March 2016 with $\alpha = 1.5$ can be seen in figure 7-4.

There are several ways of presenting uncertainty, figure 7-5 shows the mean, lower and upper limits of posterior predicted distributions. It can be said that higher uncertainty is associated with a combination of sparsity of monitoring data and higher concentrations. Although the performance is acceptable in terms of time it can be limited for longer periods of analysis. The INLA approach is the most computationally intensive compared with the other models in this work. It is natural since the rest of the alternatives just generate point estimations for the predictions. One important advantage is that problems of convergence and mixing of the sampling are not observed at all when working with the INLA algorithm.

Finally, the third study consist in producing forecast 3 days ahead when fitting the model for 16 sites. The three locations that were leaved out are: BL0, GN3, ST5. This exercise considers 7 months of data to produce the result and the settings for the INLA mesh are the same than the one used for the previous examples. The results can be seen in table 7.4. In this table the column names are α, prior κ, prior τ and the RMSE for 1,2 and 3 days ahead. In term of RMSE, there are not noticeable differences in the results when different settings are considered. It is also important to say that the error increases with the number of days ahead forecasted.

7.6.1 Discussion

In this chapter, we moved from the kriging based approach to spatio-temporal modelling, as described in Chapter 6, and implemented a spatio-temporal model within a Bayesian hierarchical framework. The major addition and advantage in using this approach is the way that estimates of uncertainty can be produced for the predictions of the model, on top of the ability to perform spatial predictions compared to the DSTM approach (Chapter 5). At each prediction location, a full posterior distribution is estimated, which allows a variety of uncertainty measures to be obtained, including credible intervals, exceedance probabilities and others. The ability to perform spatial predictions allows maps of air pollution to be created (by predicting on a grid covering the study area) and, together with the ability to produce associated maps of uncertainty (e.g. Figures 7-4 and 7-5), this provides a powerful approach.

With any Bayesian analysis, is it important to assess the sensitivity of the choice of priors; in this case there did not seem to be any striking differences observed when running the model with different choices for the (hyper-)parameters. In this implementation, there is also a question of whether the choice of mesh will have an effect. As part of the background analyses to the results presented in this chapter, different mesh configurations were used for the LAQN data, with similar results observed in terms of the predictions and uncertainty.

			7 Months		3 Months		2 Months	
	Site	Real	$\alpha = 2$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1.5$	alpha =2	$\alpha = 1.5$
	BL0	19.1875	18.9597	19.7916	18.8985	18.8983	20.0823	20.5791
	BX1	15.7917	20.9243	20.8148	20.8863	20.8863	20.8012	20.7679
	CD1	21.3875	25.4617	25.4209	25.3545	25.3549	25.3166	24.6910
	CD9	26.4237	31.4972	30.8806	31.2702	31.2721	30.4068	29.1191
	CT3	22.7000	22.1721	22.9343	22.0230	22.0239	23.0949	23.6522
	GN0	19.8932	23.5543	23.6876	23.4768	23.4772	23.6556	23.5176
	GN2	20.6132	21.0107	22.5160	20.8108	20.8121	22.7782	24.7888
	GN3	20.8812	27.0100	26.6293	26.9824	26.9826	26.4919	25.9013
	GR4	16.2870	21.1897	21.3491	21.1275	21.1276	21.3835	21.3031
15/03/16	GR9	20.7735	26.5738	26.7515	26.4540	26.4546	26.6726	26.2816
	HG1	9.2000	7.2582	7.3054	7.2642	7.2642	7.5475	7.7932
	HR1	11.4000	9.3647	9.3763	9.3630	9.3630	9.6376	9.8632
	HV1	15.2052	20.2507	20.1204	20.2293	20.2293	20.0330	19.8660
	KC1	19.0875	21.9000	21.9914	21.8283	21.8284	21.8369	21.7805
	KC2	22.6098	28.4092	27.1070	28.5068	28.5059	26.4800	25.2647
	LH0	18.2871	23.0094	23.0926	22.8809	22.8812	22.9180	22.7846
	LW2	22.4770	28.2115	27.6335	28.2343	28.2341	27.3646	26.7395
	ST5	20.2300	23.4555	23.2064	23.5325	23.5318	22.6438	22.8453
	TH4	11.4372	11.0168	10.1863	11.1612	11.1602	10.1707	9.6719
	RMSE		4.0943	3.9391	4.0535	4.0536	3.7899	3.5966
	BL0	15.7714	18.9472	19.7656	18.8738	18.8737	20.0378	20.4831
	BX1	16.1250	20.9079	20.7849	20.8554	20.8556	20.7518	20.6696
	CD1	16.8542	25.4354	25.3718	25.3081	25.3087	25.2318	24.5331
	CD9	23.6964	31.4560	30.8062	31.2014	31.2035	30.2776	28.8871
	CT3	18.7500	22.1530	22.8958	21.9882	21.9893	23.0278	23.5103
	GN0	18.7205	23.5323	23.6460	23.4372	23.4376	23.5843	23.3781
	GN2	19.2868	20.9941	22.4791	20.7800	20.7814	22.7132	24.6282
	GN3	18.8502	26.9801	26.5748	26.9301	26.9305	26.3972	25.7236
	GR4	15.7458	21.1727	21.3171	21.0957	21.0960	21.3298	21.1971
16/03/16	GR9	19.4564	26.5448	26.6964	26.4035	26.4043	26.5763	26.0976
	HG1	9.2000	7.2624	7.3139	7.2673	7.2673	7.5677	7.8305
	HR1	16.0500	9.3670	9.3813	9.3627	9.3627	9.6517	9.8877
	HV1	15.3946	20.2357	20.0932	20.2006	20.2007	19.9892	19.7807
	KC1	15.4792	21.8816	21.9568	21.7942	21.7945	21.7798	21.6674
	KC2	19.2250	28.3759	27.0504	28.4488	28.4482	26.3858	25.0978
	LH0	15.3568	22.9885	23.0535	22.8432	22.8437	22.8526	22.6564
	LW2	19.4649	28.1786	27.5744	28.1773	28.1773	27.2624	26.5479
	ST5	17.9700	23.4338	23.1669	23.4927	23.4922	22.5806	22.7161
	TH4	11.4372	11.0173	10.1897	11.1574	11.1564	10.1831	9.6981
	RMSE		6.1174	5.9930	6.0523	6.0526	5.8090	5.5421
	BL0	32.8500	18.9347	19.7396	18.8492	18.8491	19.9936	20.3886
	BX1	26.4250	20.8915	20.7552	20.8246	20.8249	20.7029	20.5729
	CD1	30.9917	25.4091	25.3230	25.2619	25.2626	25.1478	24.3779
	CD9	34.9552	31.4150	30.7322	31.1328	31.1352	30.1496	28.6594
	CT3	28.3100	22.1340	22.8575	21.9536	21.9548	22.9613	23.3709
	GN0	28.3283	23.5103	23.6046	23.3976	23.3982	23.5135	23.2409
	GN2	31.9742	20.9775	22.4425	20.7493	20.7508	22.6489	24.4704
	GN3	30.1591	26.9501	26.5206	26.8781	26.8785	26.3034	25.5490
	GR4	25.6125	21.1558	21.2852	21.0641	21.0645	21.2765	21.0927
17/03/16	GR9	30.7078	26.5159	26.6416	26.3533	26.3543	26.4808	25.9168
	HG1	9.2000	7.2666	7.3223	7.2704	7.2704	7.5880	7.8676
	HR1	28.4083	9.3694	9.3864	9.3624	9.3624	9.6657	9.9121
	HV1	24.9809	20.2208	20.0661	20.1720	20.1722	19.9458	19.6966
	KC1	27.7333	21.8632	21.9223	21.7603	21.7607	21.7232	21.5562
	KC2	19.2250	28.3427	26.9941	28.3911	28.3906	26.2924	24.9336
	LH0	29.1291	22.9678	23.0146	22.8058	22.8063	22.7877	22.5304
	LW2	30.7497	28.1458	27.5157	28.1205	28.1208	27.1612	26.3596
	ST5	27.4300	23.4121	23.1275	23.4532	23.4527	22.5178	22.5890
	TH4	11.4372	11.0178	10.1932	11.1536	11.1527	10.1955	9.7241
	RMSE		7.4886	7.2437	7.5674	7.5670	7.2170	7.1667

Table 7.1: Evaluation of applying the INLA model to data from 01/08/15 to 14/03/16 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using $\alpha = 2$ and $\alpha = 1.5$ (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 15, 16 and 17 of March, 2016. All comparisons were made in the original scale ($\mu g/m^3$).

			7 Months		3 Months		2 Months	
	Site	Real	$\alpha = 2$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1.5$
	BL0	6.6083	8.4641	8.7325	8.6381	8.8960	9.0855	9.1347
	BX1	4.6583	7.0369	6.8852	6.9521	6.7875	6.6055	6.4742
	CD1	8.2833	10.2746	10.1528	10.2441	10.1068	10.0960	9.8428
	CD9	9.3778	13.4624	12.9312	13.1968	12.6537	12.4934	11.8533
	CT3	7.05	9.5642	9.3779	9.5440	9.3390	9.3840	9.1022
	GN0	11.5432	10.1150	10.1228	10.1095	10.1097	10.1424	10.1093
	GN2	6.4976	10.1529	10.2802	10.1995	10.2822	10.3511	10.2277
	GN3	10.6922	10.5591	10.4601	10.5004	10.4135	10.3946	10.4102
	GR4	6.5909	7.5719	7.5108	7.5476	7.4867	7.4238	7.3502
26/03/2016	GR9	8.0627	9.7381	9.5129	9.6617	9.4252	9.3282	9.0855
	HG1	9.2	6.2065	6.4544	6.3220	6.6187	6.9928	7.3771
	HR1	6.6458	8.0799	8.0067	8.0172	7.9260	7.8124	7.7380
	HV1	7.724	7.9469	7.8875	7.8899	7.8296	7.7545	7.7354
	KC1	6.025	8.0050	7.9107	7.9755	7.8638	7.8066	7.6318
	KC2	19.225	11.7236	11.8890	11.7282	11.9274	11.9908	12.5278
	LH0	6.6806	6.9639	6.9249	6.8656	6.8314	6.6875	6.6935
	LW2	14.2416	14.0695	14.5321	14.1558	14.6589	14.8498	15.3275
	ST5	9.36	10.0329	10.0432	10.0197	10.0160	10.0320	9.9807
	TH4	11.4372	7.3338	7.5254	7.4110	7.6628	7.8929	8.3349
	RMSE		2.7522	2.6462	2.7163	2.5955	2.5411	2.3313
	BL0	3.5455	8.4638	8.7321	8.6360	8.8929	9.0719	9.1190
	BX1	1.5696	7.0380	6.8887	6.9545	6.7954	6.6147	6.5038
	CD1	4.6833	10.2720	10.1489	10.2371	10.0964	10.0718	9.8129
	CD9	6.0587	13.4551	12.9192	13.1791	12.6252	12.4400	11.7779
	CT3	5.23	9.5625	9.3760	9.5392	9.3335	9.3676	9.0874
	GN0	7.0242	10.1126	10.1190	10.1029	10.0992	10.1175	10.0736
	GN2	3.5175	10.1505	10.2760	10.1926	10.2706	10.3239	10.1894
	GN3	6.5583	10.5562	10.4554	10.4925	10.4010	10.3669	10.3678
	GR4	2.8957	7.5725	7.5131	7.5485	7.4913	7.4263	7.3667
27/03/2016	GR9	4.3177	9.7362	9.5107	9.6567	9.4191	9.3125	9.0711
	HG1	9.2	6.2083	6.4586	6.3256	6.6271	6.9986	7.3927
	HR1	4.0167	8.0800	8.0079	8.0169	7.9285	7.8116	7.7482
	HV1	3.305	7.9471	7.8890	7.8899	7.8325	7.7541	7.7456
	KC1	3.3167	8.0051	7.9121	7.9753	7.8666	7.8058	7.6438
	KC2	19.225	11.7190	11.8803	11.7159	11.9042	11.9434	12.4345
	LH0	3.6658	6.9651	6.9282	6.8682	6.8391	6.6960	6.7200
	LW2	7.4832	14.0611	14.5148	14.1340	14.6136	14.7628	15.1580
	ST5	6.11	10.0306	10.0396	10.0134	10.0061	10.0084	9.9477
	TH4	11.4372	7.3347	7.5276	7.4122	7.6664	7.8910	8.3341
	RMSE		5.0848	5.0178	5.0471	4.9692	4.9253	4.7824
	BL0	3.2136	8.4634	8.7317	8.6338	8.8898	9.0584	9.1035
	BX1	2.2177	7.0392	6.8921	6.9569	6.8033	6.6239	6.5330
	CD1	4.487	10.2694	10.1451	10.2301	10.0860	10.0478	9.7837
	CD9	6.6365	13.4477	12.9073	13.1615	12.5969	12.3874	11.7044
	CT3	3.71	9.5609	9.3741	9.5345	9.3279	9.3513	9.0729
	GN0	5.5313	10.1103	10.1152	10.0964	10.0888	10.0929	10.0386
	GN2	2.7521	10.1480	10.2718	10.1857	10.2590	10.2970	10.1520
	GN3	5.9556	10.5532	10.4508	10.4847	10.3886	10.3395	10.3264
	GR4	3.0714	7.5731	7.5153	7.5494	7.4959	7.4287	7.3829
28/03/2016	GR9	3.8921	9.7343	9.5085	9.6516	9.4131	9.2969	9.0570
	HG1	9.2	6.2102	6.4627	6.3292	6.6355	7.0043	7.4080
	HR1	3.2217	8.0800	8.0092	8.0166	7.9309	7.8107	7.7583
	HV1	3.9784	7.9473	7.8905	7.8899	7.8354	7.7537	7.7556
	KC1	3.0136	8.0053	7.9136	7.9751	7.8693	7.8050	7.6556
	KC2	19.225	11.7143	11.8715	11.7038	11.8811	11.8967	12.3436
	LH0	2.1352	6.9663	6.9316	6.8707	6.8468	6.7044	6.7460
	LW2	9.7832	14.0527	14.4975	14.1122	14.5688	14.6772	14.9936
	ST5	6.8557	10.0283	10.0360	10.0071	9.9963	9.9851	9.9155
	TH4	11.4372	7.3355	7.5297	7.4133	7.6700	7.8890	8.3333
	RMSE		5.2020	5.1257	5.1614	5.0722	5.0187	4.8646

Table 7.2: Evaluation of applying the INLA model to data from 12/08/15 to 25/03/16 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using $\alpha = 2$ and $\alpha = 1.5$ (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 26, 27 and 28 of March, 2016. All comparisons were made in the original scale ($\mu g/m^3$).

			7 Months		3 Months		2 Months	
	Site	Real	$\alpha = 2$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1.5$
	BL0	6.3750	9.4194	9.8693	9.0929	9.5827	9.3423	9.8470
	BX1	5.2333	10.2037	10.3492	10.2488	10.3877	10.2564	10.4083
	CD1	10.8458	14.6430	14.5875	14.7074	14.6599	14.6801	14.6279
	CD9	10.6696	20.2102	19.3781	20.8249	19.9507	20.4324	19.4927
	CT3	14.5543	12.4857	12.3998	12.5598	12.4290	12.5084	12.4200
	GN0	11.0350	13.4091	13.5885	13.3962	13.5875	13.4437	13.6486
	GN2	6.8770	12.9683	12.9225	13.0416	12.9988	13.0000	12.9514
	GN3	10.2575	16.5829	16.3973	16.8030	16.6371	16.6608	16.4495
	GR4	5.7458	10.8869	10.7928	11.0075	10.9225	10.9102	10.8006
14/10/2015	GR9	6.8466	13.5555	13.3148	13.7627	13.5539	13.5987	13.3242
	HG1	12.1833	16.2483	16.3578	16.0675	16.1949	16.2705	16.4187
	HR1	3.9083	9.0726	9.0339	9.2019	9.1644	9.1261	9.0771
	HV1	5.3969	9.4033	9.2937	9.6193	9.5083	9.4621	9.3256
	KC1	5.5000	10.7799	10.7557	10.8483	10.8516	10.8097	10.7768
	KC2	16.3481	13.4125	13.3488	13.5790	13.4580	13.4697	13.4011
	LH0	4.9083	9.6627	9.6587	9.9301	9.9114	9.7719	9.7367
	LW2	12.0641	16.7332	16.9947	16.6634	16.9220	16.7895	17.0865
	ST5	11.7100	14.7800	14.7260	14.9602	14.9242	14.8662	14.7988
	TH4	11.4372	9.5787	9.5786	9.6071	9.5696	9.5851	9.5878
	RMSE		4.8716	4.7896	4.9947	4.9129	4.9263	4.8299
	BL0	9.7250	9.4189	9.8676	9.0974	9.5911	9.3433	9.8480
	BX1	8.7542	10.2009	10.3451	10.2523	10.3947	10.2560	10.4077
	CD1	15.7458	14.6241	14.5580	14.7060	14.6572	14.6706	14.6112
	CD9	16.7136	20.1660	19.3131	20.8142	19.9318	20.4080	19.4529
	CT3	14.5543	12.4752	12.3844	12.5610	12.4319	12.5037	12.4123
	GN0	14.2139	13.3951	13.5657	13.3964	13.5875	13.4370	13.6360
	GN2	11.3665	12.9560	12.9039	13.0422	13.0003	12.9943	12.9416
	GN3	13.6809	16.5557	16.3551	16.7988	16.6288	16.6465	16.4246
	GR4	9.6750	10.8819	10.7865	11.0103	10.9286	10.9086	10.7987
15/10/2015	GR9	10.9183	13.5409	13.2937	13.7625	13.5541	13.5917	13.3130
	HG1	15.6625	16.2226	16.3159	16.0643	16.1879	16.2572	16.3939
	HR1	9.2792	9.0731	9.0362	9.2062	9.1735	9.1275	9.0805
	HV1	8.4433	9.4029	9.2948	9.6234	9.5169	9.4630	9.3284
	KC1	9.7583	10.7753	10.7495	10.8513	10.8577	10.8083	10.7750
	KC2	16.3766	13.3985	13.3275	13.5790	13.4584	13.4630	13.3895
	LH0	9.4458	9.6615	9.6581	9.9339	9.9193	9.7723	9.7382
	LW2	19.0288	16.7053	16.9481	16.6593	16.9128	16.7748	17.0586
	ST5	13.4300	14.7605	14.6957	14.9585	14.9208	14.8563	14.7814
	TH4	11.4372	9.5778	9.5784	9.6111	9.5781	9.5858	9.5897
	RMSE		1.7967	1.6728	1.9151	1.7855	1.8323	1.6857
	BL0	20.6917	9.4184	9.8660	9.1018	9.5994	9.3443	9.8490
	BX1	18.7625	10.1981	10.3411	10.2558	10.4018	10.2555	10.4070
	CD1	24.9708	14.6053	14.5288	14.7046	14.6545	14.6612	14.5946
	CD9	24.1276	20.1219	19.2486	20.8035	19.9130	20.3837	19.4134
	CT3	14.5543	12.4647	12.3690	12.5622	12.4347	12.4991	12.4046
	GN0	21.1872	13.3812	13.5430	13.3966	13.5876	13.4303	13.6234
	GN2	22.0827	12.9437	12.8854	13.0429	13.0019	12.9886	12.9319
	GN3	24.6283	16.5287	16.3133	16.7945	16.6206	16.6323	16.3999
	GR4	19.5625	10.8769	10.7802	11.0132	10.9346	10.9071	10.7969
16/10/2015	GR9	20.7506	13.5265	13.2728	13.7623	13.5543	13.5848	13.3019
	HG1	25.6625	16.1970	16.2744	16.0610	16.1808	16.2438	16.3692
	HR1	18.1167	9.0736	9.0385	9.2106	9.1826	9.1289	9.0840
	HV1	18.6330	9.4025	9.2959	9.6274	9.5254	9.4640	9.3311
	KC1	19.4458	10.7707	10.7434	10.8543	10.8639	10.8069	10.7732
	KC2	24.0530	13.3846	13.3063	13.5790	13.4589	13.4563	13.3780
	LH0	17.0958	9.6603	9.6575	9.9377	9.9272	9.7727	9.7397
	LW2	26.6666	16.6776	16.9019	16.6552	16.9036	16.7602	17.0307
	ST5	24.2300	14.7411	14.6655	14.9568	14.9174	14.8465	14.7641
	TH4	11.4372	9.5769	9.5781	9.6152	9.5865	9.5865	9.5916
	RMSE		8.4610	8.4776	8.3657	8.3590	8.4057	8.4183

Table 7.3: Evaluation of applying the INLA model to data from 01/03/15 to 13/10/15 considering $PM_{2.5}$ data. Results are shown using 7, 3 and 2 months data when fitting the model, using $\alpha = 2$ and $\alpha = 1.5$ (see text for details). Observations are given together with predictions from the models for 19 sites in London, together with overall root mean squared error (RMSE). Evaluations are presented for 1, 2 and 3 day forecasts covering 14, 15 and 16 of October, 2015. All comparisons were made in the original scale ($\mu g/m^3$).

			RMSE		
α	prior κ	prior τ	1 day-ahead	2 days-ahead	3 days-ahead
2	default	default	1.3322	1.4492	1.3520
1.5	default	default	1.1761	1.2549	1.5390
0.9	0.01	0.1	1.0900	1.1341	1.8081
0.9	0.5	0.1	1.0900	1.1342	1.8078
0.9	0.01	0.5	1.0910	1.1362	1.7990
0.9	0.5	0.5	1.0899	1.1340	1.8087
0.5	0.01	0.1	1.0898	1.1338	1.8093
0.5	0.5	0.1	1.0915	1.1373	1.7944
0.5	0.01	0.5	1.0900	1.1342	1.8078
0.5	0.5	0.5	1.0899	1.1339	1.8088
0.1	0.01	0.1	1.0898	1.1337	1.8099
0.1	0.5	0.1	1.0896	1.1334	1.8113
0.1	0.01	0.5	1.0899	1.1340	1.8082
0.1	0.5	0.5	1.0899	1.1339	1.8090

Table 7.4: Evaluation of applying the INLA model to data from 01/08/15 to 14/03/16 considering $PM_{2.5}$ data and leaving out three sites (BL0, GN3, ST5) when fitting. Results are shown using 7 months data when fitting the model, using $\alpha = 2, 1.5, 0.9, 0.5, 0.1$ and different values for prior κ and τ (see text for details). Overall root mean squared error (RMSE) is given from the models for 3 sites in London (BL0, GN3, ST5). Evaluations are presented for 1, 2 and 3 day forecasts covering 15, 16 and 17 of March, 2016. All comparisons were made in the original scale ($\mu g/m^3$).

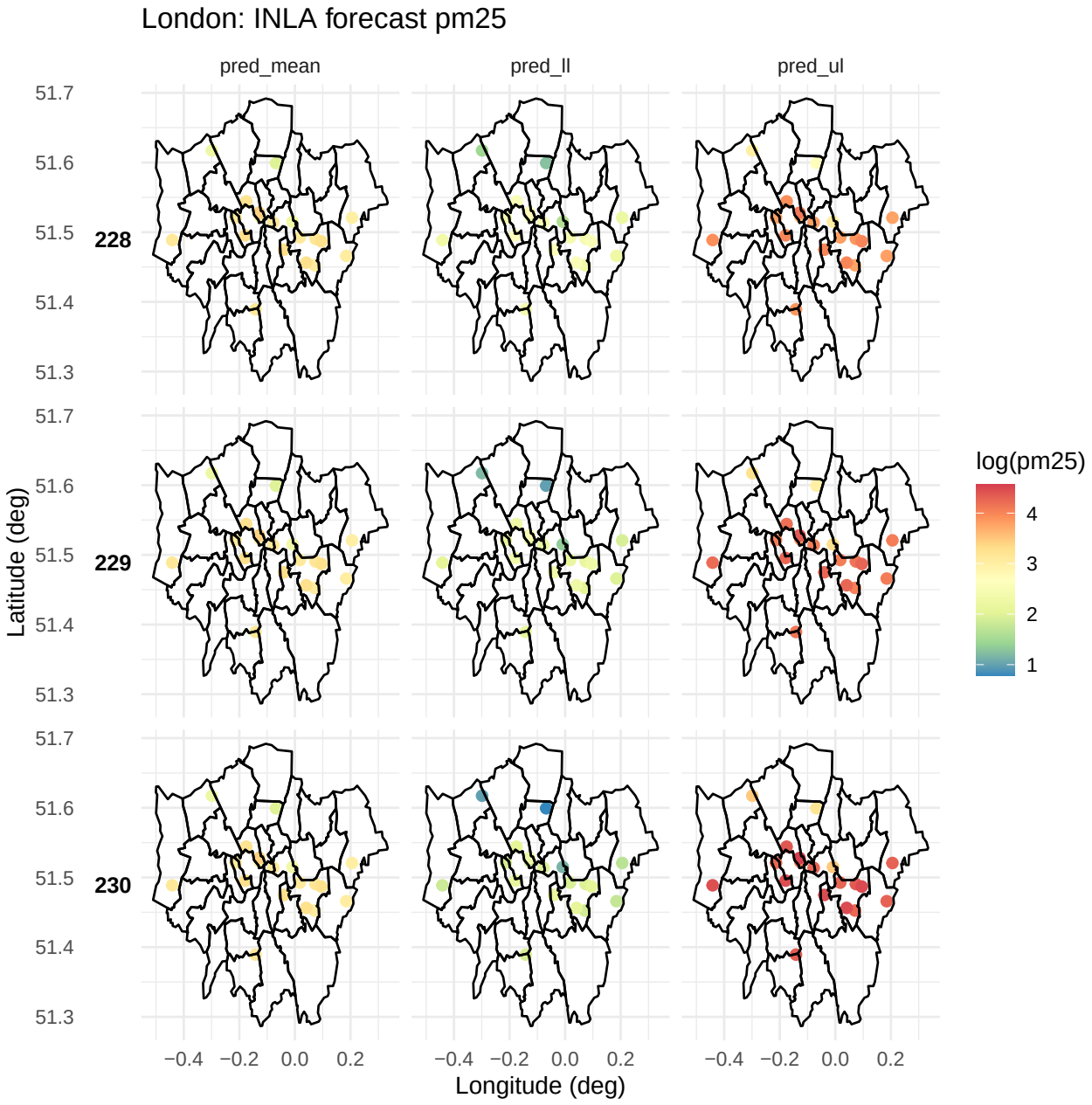

Figure 7-3: Predicted values of $PM_{2.5}$ concentrations in Greater London 15, 16 and 17 March 2016 (days: 228, 229, and 230 respectively). Values are mean, lower and upper limits of posterior predicted distributions on the logarithmic scale from a SPDE–INLA model. The period of time used when fitting the mode is from 01 August 2015 to 14 March 2016 and with $\alpha = 1.5$.

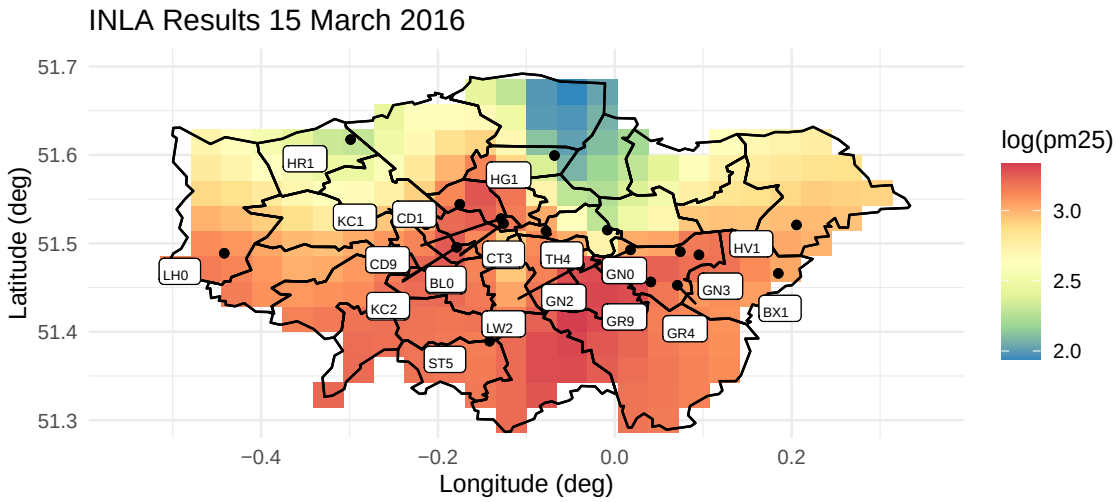

Figure 7-4: Map of predicted values of $PM_{2.5}$ concentrations in Greater London (15 March 2016). Values are means of posterior predicted distributions on the log-scale from a SPDE–INLA model. The period of time used when fitting the mode is from 01 August 2015 to 14 March 2016 and with $\alpha = 1.5$.

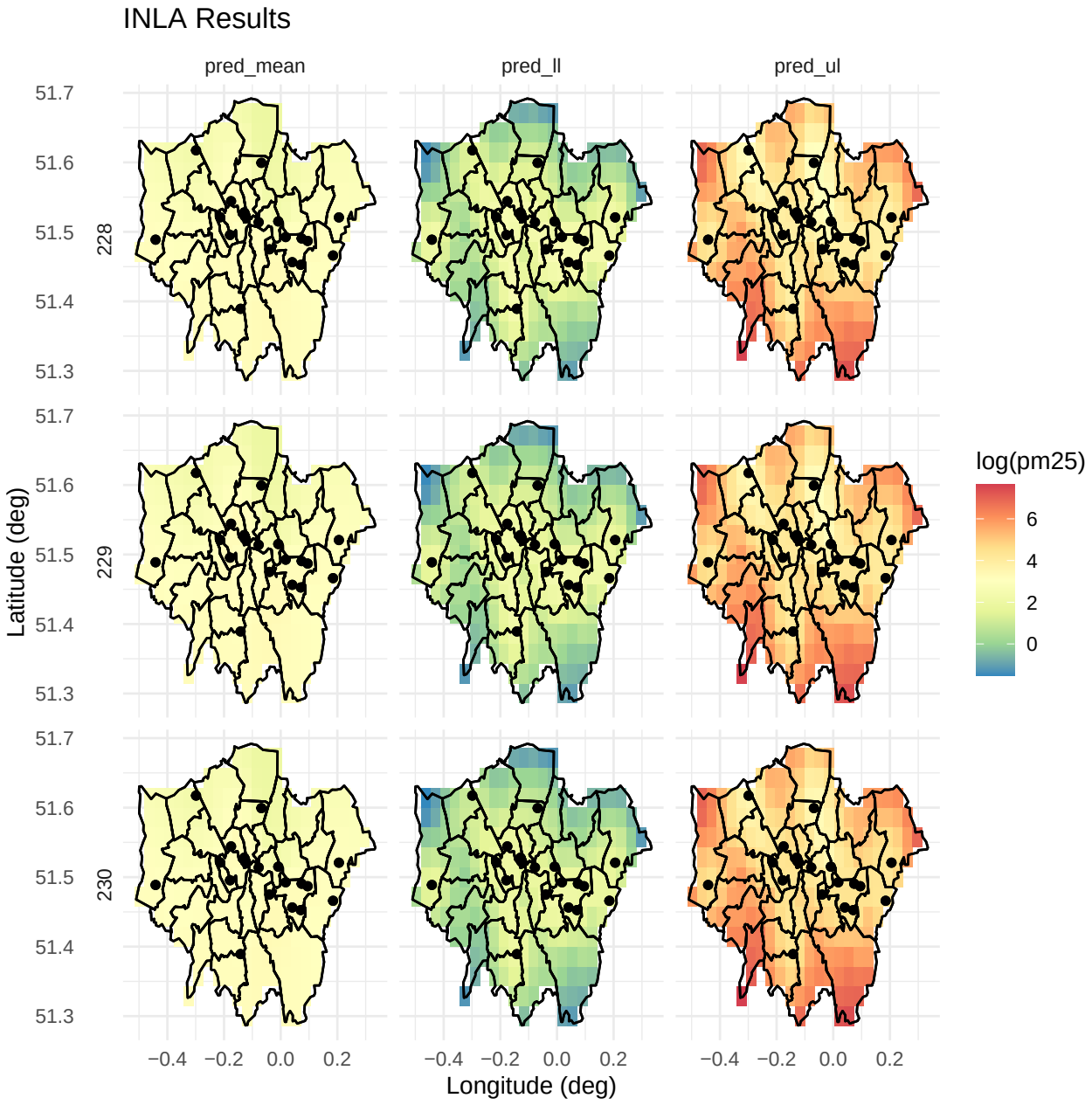

Figure 7-5: Map of predicted values of $PM_{2.5}$ concentrations in Greater London 15, 16 and 17 March 2016 (days: 228, 229, and 230 respectively). Values are mean, lower and upper limits of posterior predicted distributions on the logarithmic scale from a SPDE–INLA model. The period of time used when fitting the model is from 01 August 2015 to 14 March 2016 and with $\alpha = 1.5$.

Chapter 8

Discussion

The aim of this thesis was to develop and compare different prediction models for daily concentrations of $PM_{2.5}$. Two different approaches were developed and used to produce predictions of air pollution levels. These were Dynamic Space-Time Models (DSTM) and Bayesian Hierarchical Spatio-Temporal Models (BHSTM). We also investigated the use of spatio-temporal kriging and fixed rank kriging and their ability to produce maps of air pollution based on measurements at a distinct number of locations. Throughout the thesis, we present results of implementing the different approaches, comparing their ability to model spatio-temporal patterns in the data and to produce short-term forecasts of $PM_{2.5}$. The predictions from each of the models were compared with actual data in order to evaluate them.

8.1 Dynamic Spatio-temporal models

The Dynamic Spatio-Temporal Modelling approach presented in chapter 5 allowed future concentrations of $PM_{2.5}$ to be forecasted in a computationally efficient manner. Under this model, it is possible to include more information from other pollutants (PM_{10}, NO_2) that are closely related to $PM_{2.5}$ in urban areas to inform the predictions of $PM_{2.5}$. The main objective, when a DSTM is fitted, is to make inference on the unobserved states, which are the underlying processes from which measurements are made. An adequate DSTM representation has to be chosen, then the estimation is done by computing the conditional distributions of the quantities of interest. It is done considering given the available information and using the Kalman filter or smoother. Details of estimation for linear DSTMs with consideration of methods for dimension reduction are also explained. Two methods for inference, such as methods of moments

and EM algorithm, are implemented using the London air quality data.

In summary, the results show that the reduced dimension linear DSTM had a good performance at forecasting PM2.5 concentrations. The results of implementing the method of moments and EM algorithm results can be seen in tables 5.1 and 5.2 with overall RMSE values of as small as ca. 1.5 μg^{-3} for one day forecasts and ca. 1.6 μg^{-3} for two-day forecasts, although for some of the implementations (e.g. using different amounts of modelling data and different methods) is was closer to 2.2. Errors for three day ahead forecasts were seen to be considerably larger. An advantage of this approach is the relatively simple extension for the multivariate case that includes PM10 and NO2 concentrations. It is notable that both models have the lowest RMSE, see table 8.1 for the prediction for 15 March 2016.

In the form presented here, is not possible to include uncertainty analysis in the DSTM EM algorithm. If one is interested in uncertainty, nonparametric bootstrapping or a fully Bayesian framework could be considered as an alternative. However, it can be challenging to take bootstrap samples that adequately represent the dependence structure in the spatio-temporal data. If one is interested in a solution that includes a credible interval for the prediction, the BHM approach is an option, see Berliner *et al.* (2000) for further discussion.

8.2 Bayesian Hierarchical Spatio-Temporal Modelling

In chapter 7, a Bayesian approach for a spatio-temporal hierarchical model was proposed. It considers a temporal process, characterised by first order autoregressive dynamics and spatially correlated innovations based on a Gaussian Field. In particular, $PM_{2.5}$ concentrations were analysed from 01 August 2015 to 14 March 2016 without 15 March. Inference was performed using INLA, which was used to get the parameter posterior estimates, together with prediction and uncertainty for 15 March 2016. The prediction is similar in accuracy to the DSTM that was presented in chapter 5, see table 8.1. In addition, the spatial prediction for 16 and 17 March produced by INLA can be seen in figure 7-5 (Chapter 7). The broad spatial patterns of concentrations seen in this figure are similar to those seen in 6-4 and 6-9 (from Chapter 6), although it should be noted that the methods used for generating the predict values over space shown in figure 7-4 and figures 6-4 and 6-9 are different. The first of these is the result of a space-time forecast for 15 March 2016, while the other two are smoothed (spatial only) maps for 15 March 2016.

One advantage of the INLA approach is that a posterior distribution is generated, not

just a point estimation. In some contexts, a probability interval could be more useful than a single point estimation. Another benefit in using the BHSTM framework is that it captures correlations via a traditional covariance matrix. Therefore, it can do spatial prediction, which as discussed is a limitation under the given implementation of the DSTM approach.

Common problems in simulated-based MCMC methods, such as convergence and mixing are not present when working with the INLA algorithm. Moreover, the SPDE approach can be extended to a wide class of spatio-temporal models, with relatively simple modifications. It is always possible to consider models with more complex hierarchical structures. For instance, a non-separable covariance function or non-stationary cases characterised by parameters varying in time can be considered. Although there are many advantages of INLA, it can also be limited by the computation burden. One can conclude that the SPDE approach, combined with the INLA algorithm, is a convenient framework for performing Bayesian inference on complex spatio-temporal GFs. Even when dealing with large data sets, BHSTM is a good option to analyse air pollution data. However, the size of the output object can get large when using an important amount of temporal data. Then, its use can be restrictive in some circumstances.

8.3 Comparison of predictive ability

In this section, the approaches DSTM and BHSTM presented in chapters 5 and 7 respectively are compared. This comparison is made in terms of their ability to produce short-term forecasts of $PM_{2.5}$ for a common time period.

In the previous chapter, the results for forecasting studies were presented. Now, a summary with all of those results are shown in table 8.1. The summary contains the RMSEs from tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 7.1, 7.2 and 7.3. In particular, three days were omitted from the original data sets to evaluate the forecast over time. Some conventions in the columns names will be followed. The column Poll considers the number of pollutants 3 ($PM_{2.5}$, PM_{10}, NO_2) and 1 ($PM_{2.5}$) used when fitting the model. The names Months and Method are the number of months and different method used for the exercise. Finally, the numbers 1, 2, 3, 4, 5, 6, 7, 8 and 9 correspond to the dates 15/March/16, 16/March/16, 17/March/16, 26/March/16, 27/March/16, 28/March/16, 14/October/2015, 15/October/15 and 16/October/15 respectively. The same convention for the dates will be used for the figure that will be presented later. The forecasts for 15, 16 and 17 March 2016 were produced when fitting a model with three different time intervals up to 14/03/2016. The forecasts for 26, 27 and 28 March

2016 were produced when fitting a model with three different time intervals up to 25/03/2016. Meanwhile, the forecasts for 14, 15 and 16 October 2015 were generated considering three periods of time up to 13/10/2015.

In the table 8.1, the smallest RMSE per predicted day ahead is highlighted in blue. For 15 March 2016 the best prediction in terms of RMSE was reached with the EM algorithm using 3 months of data as input. For 16 and 17 March, the best predictions are related to the EM-algorithm (3 months) and INLA ($\alpha = 1.5$ and 2 months) respectively. Overall, the best predictions for 26, 27 and 28 March 2016 were observed for the INLA approach considering 2 months and $\alpha = 1.5$. In general, it was observed that the errors are larger when more days ahead are predicted. Finally, the best predictions for 14, 15 October 2015 were observed for the MM, considering 7 months, 7 months and 2 months respectively.

The extent of the errors observed when using the different models can be seen in figure 8-1. In this figure, the results from table 8.1 are summarised. It is noted that there is not a noticeable difference in terms of RMSE between the models INLA, with $\alpha = 2$ and $\alpha = 1.5$ for most of the cases. One advantage of the DSTM approach is the ability to incorporate information on more pollutants (PM10, NO2) when forecasting $PM_{2.5}$ in a computationally efficient manner. However, in the examples shown here incorporating multi-pollutant data did not necessarily improve the forecast performance. In general, for the multipollutant and $PM_{2.5}$

8.4 Summary

In summary, the ability to jointly model in time and space, and across pollutants, allows information to be borrowed in a variety of ways to better inform prediction. However, there may be challenges in implementing increasingly complex models, with the increasing amount of air quality related data that is becoming available.

The fundamental difference in the two approaches considered here is in how spatial dependence is incorporated within the models. In the first approach (DSTM), dependencies between data from different locations is incorporated within the transition matrix that governs how the underlying spatio-temporal process (the underlying pollution field) transitions from one time to the next. In the second approach (BHSTM), spatial dependence is explicitly modelled which enables such prediction (at unmeasured locations) but does require knowledge of the nature of the dependence which, if not known entirely (which will usually be the case), will mean making assumptions that simplify the process.

Poll	Months	Method	RMSE								
			1	2	3	4	5	6	7	8	9
	7	MM	1.5484	2.3497	15.6442	2.4365	5.2986	5.8792	4.5915	1.6364	9.5196
	7	EM	1.6178	2.2094	13.1362	2.5088	5.5134	5.8789	4.6862	1.6407	9.1590
	3	MM	1.3984	2.3549	16.2087	2.5312	5.9403	6.7702	5.3756	2.0530	9.0333
3	3	EM	1.7021	1.5920	12.8087	2.6296	5.7956	6.1452	5.3441	2.0327	8.6508
	2	MM	2.4810	3.0832	15.4644	2.4063	5.5322	6.5579	5.9838	2.6617	8.4599
	2	EM	1.8278	2.0658	13.4706	2.4354	5.4766	5.8922	6.1375	2.6014	8.0599
	7	MM	1.4607	2.4526	15.6304	3.0240	8.0873	8.8556	4.3389	1.5100	8.8834
	7	EM	1.3821	1.8657	11.9422	3.0702	7.0230	7.6074	4.5995	1.5287	8.6712
	7	INLA $\alpha = 2$	4.0943	6.1174	7.4886	2.7522	5.0848	5.2020	4.8716	1.7967	8.4610
	7	INLA $\alpha = 1.5$	3.9391	5.9930	7.2437	2.6462	5.0178	5.1257	4.7896	1.6728	8.4776
	3	MM	1.3233	2.4046	16.0316	2.6166	6.3133	6.7158	4.8503	1.9690	8.8459
1	3	EM	1.2844	1.8976	13.0422	2.7157	6.1362	6.4216	4.7254	1.8563	8.9519
	3	INLA $\alpha = 2$	4.0535	6.0523	7.5674	2.7163	5.0471	5.1614	4.9947	1.9151	8.3657
	3	INLA $\alpha = 1.5$	4.0536	6.0526	7.5670	2.5955	4.9692	5.0722	4.9129	1.7855	8.3590
	2	MM	1.4374	2.2020	15.5008	5.5997	15.4472	17.1799	4.3821	1.9899	7.1664
	2	EM	1.5191	2.2107	13.4923	5.1560	6.6370	10.8671	4.5588	1.7737	8.3094
	2	INLA $\alpha = 2$	3.7899	5.8090	7.2170	2.5411	4.9253	5.0187	4.9263	1.8323	8.4057
	2	INLA $\alpha = 1.5$	3.5966	5.5421	7.1667	2.3313	4.7824	4.8646	4.8299	1.6857	8.4183

Table 8.1: Evaluation of applying the DSTM and INLA models to data from three different periods of time (01/08/15-14/03/16, 12/08/15-25/03/16 and 01/03/15-13/10/15) . Results are shown using 7, 3 and 2 months data when fitting the model, using the methods of moments (MM), expectation-maximisation (EM) and INLA approaches to inference (see text for details). Evaluations are presented for 1, 2 and 3 day forecasts with the overall root mean squared error (RMSE). Here, the RMSEs were previously presented in tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 7.1, 7.2 and 7.3. All comparisons were made in the original scale (μg/m^3).

Pollutants such as $PM_{2.5}$ are complex combinations of particles and so making such assumptions can be restrictive and possibly unrealistic. However, as will all statistical modelling it may be necessary to take a pragmatic approach that balances the need to make such assumptions in order to achieve the goals of the analysis and produce useful results given the quality and quantity of the data that is available.

As mentioned before, there are some advantages of the DSTM approach. It is a flexible and general framework that allows one to analyse and forecast spatio-temporal data. Due to its extensibility, it also possible to include multi-pollutant information in the analysis. Moreover, this model is also computationally efficient in terms of time, but cannot handle missing values. One concern is that this model does not take the uncertainty of the parameter estimates into account. Another disadvantage in the implementation considered here, is that there was no spatial structure in the covariance matrix associated with the process model. Therefore, it is not possible to produce spatial predictions (maps) except by applying another method to the predictions, such as IDW or kriging techniques.

One disadvantage of the DSTM approach is that uncertainty estimates can be difficult to obtain for the parameter estimates. The issue of missing values is also a problem in the implementation of this type of models. Many of these issues are dealt with using a Bayesian approach, in which a unifying framework for the integration of spatio-

temporal modelling can be applied. This framework allows estimation of missing data with uncertainty estimates obtained from the posterior (predictive) distributions. With the model, future forecasts are set up as missing data and treated the same way, i.e. as unknown parameters for which posterior distributions are obtained. However, INLA objects can get large which can make computation using long time-series impractical. A full comparison should therefore take into account prediction capability, complexity and computational costs.

It is noted that statistical models do not usually incorporate an extensive base of scientific knowledge. However, some researchers have developed methods for incorporating outputs from numerical chemical transport models within statistical models for air pollution, for example Zidek *et al.* (2012); Sahu (2016). This idea of 'hybrid' models is an area for future research that seems particularly well suited to forecasting air quality over space and time, where deterministic air quality forecasting models could be integrated with data measurements in a BSTHM.

When dealing with larger data sets, for example daily data over longer time periods, more monitoring locations and/or the incorporation of information from numerical models, it would be useful to explore the idea of integrating dimension reduction techniques (as discussed in chapter 5) within a Bayesian modelling approach. The aim would be to reduce the computational burden of performing Bayesian inference, which although it is very efficient using INLA can result in very large INLA objects which can hinder large-scale estimation and prediction. Other avenues for future research would be to incorporate multivariate modelling to allow borrowing of information over space-time-pollutants (as was performed using a DSTM in chapter 5) within the BSTHM.

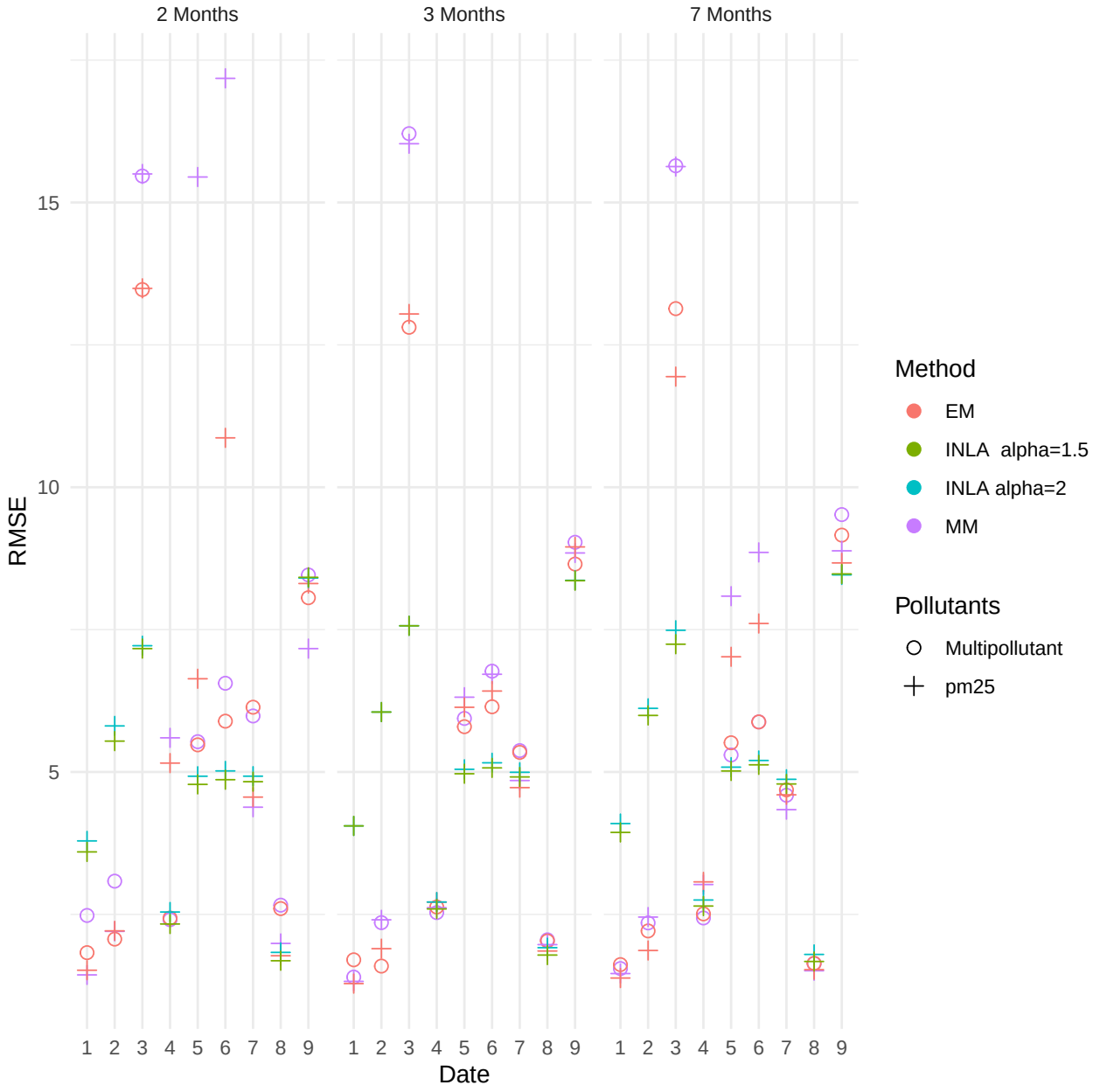

Figure 8-1: Evaluation of applying the DSTM and INLA models to data from three different periods of time (01/08/15-14/03/16, 12/08/15-25/03/16 and 01/03/15-13/10/15) . Results are shown using 2, 3 and 7 months data when fitting the model, using the methods of moments (MM), expectation-maximisation (EM) and INLA approaches to inference (see text for details). Evaluations are presented for 1, 2 and 3 day forecasts with the overall root mean squared error (RMSE). The numbers 1, 2, 3, 4, 5, 6, 7, 8 and 9 on the x-axis correspond to the dates 15/March/16, 16/March/16, 17/March/16, 26/March/16, 27/March/16, 28/March/16, 14/October/2015, 15/October/15 and 16/October/15 respectively. The values of the RMSEs can be seen in table 8.1. All comparisons were made in the original scale of the data (μg/m^3).

Appendix A

Matrix-Algebra Definitions and Properties

Kronecker Product

Definition A.0.1. Consider two matrices, an $n_a \times m_a$ matrix, $\mathbf{A}$, and an $n_b \times m_b$ matrix, $\mathbf{B}$. The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is given by the $n_a n_b \times m_a m_b$ matrix $\mathbf{A} \otimes \mathbf{B}$ defined as

$$\mathbf{A} \otimes \mathbf{B} \equiv \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1m_a}\mathbf{B} \\ \vdots & \vdots & \vdots \\ a_{n_a 1}\mathbf{B} & \cdots & a_{n_a m_a}\mathbf{B} \end{bmatrix}. \tag{A.1}$$

The Kronecker product has some properties properties that facilitate matrix representations. If $\mathbf{A}$ is $n_a \times n_a$ and $\mathbf{B}$ is $n_b \times n_b$, the inverse and determinant of the Kronecker product can be expressed in terms of the Kronecker product of the inverses and determinants of the individual matrices, respectively:

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}, \tag{A.2}$$

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^{n_a} |\mathbf{B}|^{n_b}. \tag{A.3}$$

Kronecker products are useful in the context of spatio-temporal processes. They can provide a helpful way to represent spatio-temporal covariance matrices for separable processes. Let us consider

$$\{Y(\mathbf{s}_{ij}; t_j) : i = 1, \ldots, m; j = 1, \ldots, T\},$$

and define $\mathbf{C}_y^{(\mathbf{s})}$ to be the $m \times m$ matrix of purely spatial covariances and $\mathbf{C}_y^{(t)}$ to be the $T \times T$ matrix of purely temporal covariances. Then the $mT \times mT$ spatio-temporal covariance matrix can be written as, $\mathbf{C}_y = \mathbf{C}_y^{(t)} \otimes \mathbf{C}_y^{(\mathbf{s})}$ if the process is separable. This is not realistic for many processes but it is advantageous because of the inverse property, $\mathbf{C}_y^{-1} = (\mathbf{C}_y^{(t)})^{-1} \otimes (\mathbf{C}_y^{(\mathbf{s})})^{-1}$.

Kronecker products are also useful for spatio-temporal modelling for forming spatio-temporal basis functions. Let us construct an $m \times n_{\alpha,\mathbf{s}}$ matrix $\mathbf{\Phi}$ by evaluating $n_{\alpha,\mathbf{s}}$ temporal basis functions at m spatial locations, and a $T \times n_{\alpha,t}$ matrix Ψ by evaluating $n_{\alpha,t}$ temporal basis functions at T temporal locations. Then the matrix constructed from spatio-temporal basis functions formed through the tensor product of the spatial and temporal basis functions and evaluated at all combinations of spatial and temporal locations is given by the $mT \times n_{\alpha,\mathbf{s}} n_{\alpha,t}$ matrix $\mathbf{B} = \mathbf{\Psi} \otimes \mathbf{\Phi}$. Basis functions can be used to construct spatio-temporal covariance functions. Using a set of basis functions constructed through the tensor product yields a class of spatio-temporal covariance functions that are in general not separable.

Non-negative-definite and positive-definite matrices

Definition A.0.2. Consider a $p \times p$ symmetric and real-valued matrix, $\mathbf{A}$. If, for any non-zero real-valued vector $\mathbf{x}$, the scalar given by the *quadratic form* $\mathbf{x}'\mathbf{A}\mathbf{x}$ is non-negative, it is said that $\mathbf{A}$ is a *non-negative-definite* matrix. If $\mathbf{x}'\mathbf{A}\mathbf{x}$ is strictly positive for any $\mathbf{x} \neq \mathbf{0}$, it is said that $\mathbf{A}$ is a *positive-definite matrix*.

Matrix inverse

Definition A.0.3. Let us consider the $p \times p$ square matrix, $\mathbf{A}$. If the matrix $\mathbf{B}$ such that $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{I}_p$ exists, it is known as the *inverse matrix* of $\mathbf{A}$, and it is denoted as $\mathbf{A}^{-1}$.

If the inverse matrix exists, it is said that the matrix is invertible, and it follows that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_p$. Not every square matrix has an inverse, but every positive-definite matrix is invertible. The inverse matrix is also positive-definite.

Matrix square root

Definition A.0.4. Let $\mathbf{A}$ be a $p \times p$ positive-definite matrix. Then there exists a matrix $\mathbf{B}$ such that $\mathbf{A} = \mathbf{B}\mathbf{B} = \mathbf{B}^2$, it is said that $\mathbf{B}$ is the matrix square root of $\mathbf{A}$ and it is denoted by $\mathbf{A}^{1/2}$.

The matrix square root of a positive-definite matrix is also positive-definite. The inverse matrix can be written as $\mathbf{A}^{-1} = \mathbf{A}^{-1/2}\mathbf{A}^{-1/2}$, where $\mathbf{A}^{-1/2}$ is the inverse of $\mathbf{A}^{1/2}$.

Spectral decomposition

Let $\mathbf{A}$ be a $p \times p$ symmetric matrix of real values. This matrix can be decomposed as

$$\mathbf{A} = \sum_{k=1}^{p} \lambda_k \phi_k \phi_k' = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}, \tag{A.4}$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\mathbf{\Phi} = [\phi_1, \ldots, \phi_p]$, and $\{\lambda_k\}$ are called the *eigenvalues* that are associated with the *eigenvectors*, $\{\phi_k\}$, for $k = 1, \ldots, p$. These vectors are orthogonal, i.e., $\mathbf{\Phi}\mathbf{\Phi}' = \mathbf{\Phi}'\mathbf{\Phi} = \mathbf{I}_p$. Observe that for a symmetric non-negative-definite matrix $\mathbf{A}$, then $\lambda_k \geq 0$, and for a symmetric positive-definite matrix $\mathbf{A}$, then $\lambda_k > 0$ for all $k = 1, \ldots, p$. The matrix square root and its inverse can be written as $\mathbf{A}^{1/2} = \mathbf{\Phi}\text{diag}(\lambda_1^{1/2}, \ldots, \lambda_p^{1/2})\mathbf{\Phi}'$ and $\mathbf{A}^{-1/2} = \mathbf{\Phi}\text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_p^{-1/2})\mathbf{\Phi}'$, respectively.

Eigenvalues of the Transition Matrix

Let us consider the first-order vector autoregressive model,

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \tag{A.5}$$

where $\mathbf{Y}_t$ is an n-dimensional vector and $\mathbf{M}$ is an $n \times n$ real-valued transition matrix.

The eigenvalues and eigenvectors can tell us something about the dynamical properties of the model. Taken together, an eigenvalue–eigenvector pair is sometimes referred to as an *eigenmode*. Let us assume that $\lambda_i = a_i + b_i\sqrt{-1}$ (where $b_i = 0$ if λ_i is real-valued), and define the modulus to be $|\lambda_i| = \sqrt{a_i^2 + b_i^2}$. If $\max\{|\lambda_i| : i = 1, \ldots, n\} \geq 1$ then the eigenmode, and hence the model, is unstable, and $\mathbf{Y}_t$ will grow without bound as t increases. On the other hand, if the maximum modulus of all the eigenvalues is less than 1, then the model is stable. For more details see Cressie & Wikle (2011), section 3.2.1.

Singular value decomposition (SVD)

Let $\mathbf{A}$ be a $p \times n$ matrix of real values. The matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$, where $\mathbf{U}$ and $\mathbf{V}$ are $p \times p$ and $n \times n$ orthogonal matrices, respectively. The $p \times n$ matrix $\mathbf{D}$ contains all zeros except for the (k, k)th non-negative elements, $\{d_k : k = 1, 2, \ldots, \min(p, n)\}$ which are known as *singular values*.

Bibliography

Abbey, DE, Colome, SD, Mills, PK, Burchette, R., Beeson, WL & Tian, Y. 1993. Chronic disease associated with long-term concentrations of nitrogen dioxide. *Journal of Esposure Analysis and Environmental Epidemiology;(United States)*, **3**.

Abramowitz, M. & Stegun, I., 1972. *Handbook of mathematical functions.* Dover Publications, New York.

Anderson, D.O. 1967. The effects of air contamination on health. I. *Canadian Medical Association Journal*, **97**, 528.

Anderson, H Ross. Health effects of air pollution episodes. In *Air pollution and health*, pages 461–482. Elsevier, 1999.

Anderson, H.R. 2009. Air pollution and mortality: A history. *Atmospheric Environment*, **43**, 142–152.

Ayres, S.M., Evans, R.G. & Buehler, M.E. 1972. Air pollution: A major public health problem. *Critical Reviews in Clinical Laboratory Sciences*, **3**, 1–40.

Balakrishnan, Sivaraman, Wainwright, Martin J & Yu, Bin. 2017. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, **45**, 77–120.

Berliner, L. Mark, Wikle, Christopher K. & Cressie, Noel A. 2000. Long-lead prediction of pacific ssts via bayesian dynamic modeling. *American Meteorological Society.*

Berliner, Mark. 1996. Hierarchical bayesian time series models. *Fundamental Theories of Physics*, **79**, 15–22.

Bernardo, Jose M. & Smith, Adrian F. M., 1994. *Bayesian Theory.* Wiley series in probability and mathematical statistics, New York, first edition.

Blangiardo, M. & Cameletti, M., 2015. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M. *et al.* 2003. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology*, **14**, 228–239.

Burrough, P.A., McDonnell, R.A. & McDonnell, R., 1998. *Principles of geographical information systems*. Oxford University Press. Oxford.

Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. 2013. Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, **97**, 109–131.

Cameletti, Michela, Ignaccolo, Rosaria & Bande, Stefano. 2011. Comparing spatio-temporal models forparticulate matter in piemonte. *Envirometrics*, **22**, 985–996.

Carroll, RJ, Chen, R., George, EI, Li, TH, Newton, HJ, Schmiediche, H. & Wang, N. 1997. Ozone Exposure and Population Density in Harris County, Texas. *Journal of the American Statistical Association*, **92**.

Carslaw, David. *The openair manual open-source tools for analysing air pollution data*. University of York and Ricardo Energy & Environment, 2019.

Carslaw, David C. & Ropkins, Karl. 2012. openair — an r package for air quality data analysis. *Environmental Modelling & Software*, **27–28**, 52–61. ISSN 1364-8152. doi:10.1016/j.envsoft.2011.09.008.

Casella, G. & Berger, R. L., 2002. *Statistical Inference*. Duxbury, New York, second edition.

Chew, FT, Goh, DYT, Ooi, BC, Saharom, R., Hui, JKS & Lee, BW. 1999. Association of ambient air-pollution levels with acute asthma exacerbation among children in Singapore. *Allergy*, **54**, 320–329.

Chiles, J.P. & Delfiner, P., 1999. *Geostatistics: modeling spatial uncertainty*. Wiley, New York.

Chung, Moo K. 2020. Introduction to random fields. *arXiv preprint arXiv:2007.09660*.

Collins, J. & Lewis, D. 2000. Hydrogen sulfide: evaluation of current California air quality standards with respect to protection of children. *Report of the Air Toxi-*

cology and Epidemiology Section, California Office of Environmental Health Hazard Assessment, pages 1–25.

Congdon, Peter D, 2019. *Bayesian hierarchical models: with applications using R*. CRC Press.

Cressie, Noel & Huang, Hsin-Cheng. 1999. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Associatione*, **94**, 1330–1340.

Cressie, Noel & Johannesson, Gardar. 2008. Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B*, **70**, 209–226.

Cressie, Noel, Shi, Tao & Kang, Emily L. 2010. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, **19**, 724–745. URL https://doi.org/10.1198/jcgs.2010.09051.

Cressie, Noel & Wikle, Christopher K., 2019. *Spatio-temporal statistics with R*. CRC Press, first edition.

Cressie, Noel A. C., 1993. *Statistics for Spatial Data*. John Wiley & Sons., New York.

Cressie, Noel A. C. & Wikle, Christopher K., 2011. *Statistics for Spatio-Temporal Data*. John Wiley & Sons., New Jersey.

Davies, D. S. C. Annual report of the chief medical officer 2017 . health impacts of all pollution. what do we know?, 2017. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/690846/CMO_Annual_Report_2017_Health_Impacts_of_All_Pollution_what_do_we_know.pdf.

Demidenko, Eugene, 2013. *Mixed Models: Theory and Applications with R*. Wiley Series in Probability and Statistics, second edition.

Diggle, P.J. & Ribeiro, P.J., 2007. *Model Based Geostatistics*. Springer.

Dominici, F., Daniels, M., Zeger, S. & Samet, J. 2002. Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, **97**, 100–111.

Dominici, Francesca, Samet, J. M. & Zeger, S. L. 2000. Combining evidence on air pollution and daily mortality from the 20 largest us cities: A hierarchical modelling strategy. *Environmental Health Perspectives*, **163**, 263–302.

Ecker, M.D. & Gelfand, A.E. 1997. Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 347–369.

European Commission. Standards - Air Quality - Environment - European Commission, 2017. URL http://ec.europa.eu/environment/air/quality/standards.htm.

Gamerman, D. & Lopes, H.F., 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall/CRC.

Gelfand, Alan E. & Smith, Adrian F. M. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, Andrew. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki & Rubin, Donald B, 2013. *Bayesian data analysis*. CRC press.

(GLA), Greater London Authority. London datastore, 2021. URL https://data.london.gov.uk/.

Gräler, Benedikt, Pebesma, Edzer & Heuvelink, Gerard. 2016. Spatio-temporal interpolation using gstat. *The R Journal*, **8**, 204–218. URL https://journal.r-project.org/archive/2016/RJ-2016-014/index.html.

Guerreiro, C. B. B., Foltescu, V. & De Leeuw, F. 2014. Air quality status and trends in Europe. *Atmospheric Environment*, **98**, 376–384.

Guttorp, P., Meiring, W. & Sampson, P.D. 1994. A space-time analysis of ground-level ozone data. *Environmetrics*, **5**, 241–254.

Handcock, M.S. & Stein, M.L. 1993. A bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Heimann, H. 1961. Effects of air pollution on human health. *Air pollution*, pages 159–220.

Houseago-Stokes, Richenda E. & Challenor, Peter G. 2004. Using PPCA to Estimate EOFs in the Presence of Missing Values. *Journal of Atmospheric and Oceanic Technology*, **21**, 1471–1480.

Huerta, G, Sansó, B & Stroud, J. R. 2004. A spatiotemporal model for mexico city ozone levels. *J. Royal Statistical Society*, **53**, 231–248.

Janssena, S., Dumontb, G., Fierensb, F. & Mensink, C. 2008. Spatial interpolation of air pollution measurements using corine land cover data. *Atmospheric Environment*, **42(20)**, 4884–4903.

Jones, David R. Bayes and empirical bayes methods for date analysis. bp carlin, ta louis, 2nd edn. boco raton: Chapman and hall/crc press, 2000, pp. 418,£ 34.99. isbn: 1-584-88170-4., 2001.

Kalnay, Eugenia, 2003. *Atmospheric modeling, data assimilation, and predictability.* Cambridge university press.

Kang, Emily L. & Cressie, Noel. 2011. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association*, **106**, 972–983.

Katzfuss, M. 2017. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association,*, **112**, 201–214.

Kolehmainen, M., Martikainen, H. & Ruuskanen, J. 2001. Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment*, **35**, 815–825.

Krainski, E. T., Gómez-Rubio, V., Bakka H, Lenzi A., D., Castro-Camilo, D., Simpson, F., Lindgren & H., Rue, 2018. *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. CRC Press.

Kuklinska, K., Wolska, L. & Namiesnik, J. 2015. Air quality policy in the US and the EU–a review. *Atmospheric Pollution Research*, **6**, 129–137.

Kyriakidis, P.C. & Journel, A.G. 1999. Geostatistical space-time models: a review. *Mathematical Geology*, **31(6)**, 651–684.

Laden, F., Neas, L. M., Dockery, D. W. & Schwartz, J. 2000. Association of fine particulate matter from different sources with daily mortality in six u.s cities. *Enviromental Health Prospectives*, **108**, 941–947.

Le, N.D. & Zidek, J.Z. 1992. Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal Multivariate Analysis*, **43**, 351–374.

Le, Nhu D., Sun, Weimin & Zidek, James V. 1997. Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**, 501–510.

Le, Nhu D. & Zidek, James V., 2006. *Statistical Analysis of Environmental Space-Time Processes*. Springer, New York.

Lee, Duncan & Shaddick, Gavin. 2010. Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics*, **66**, 1238–1246.

Li, Lixin, Losser, Travis, Yorke, Charles & Piltner, Reinhard. 2014. Fast inverse distance weighting-based spatiotemporal interpolation: A web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the contiguous u.s. using parallel programming and k-d tree. *International Journal of Environmental Research and Public Health*, **11**, 9101–9141. doi:10.3390/ijerph110909101.

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

London Air Quality Network. London air quality network – summary report. Technical report, Environmental Research Group (King's College London), 2018.

Lunn, David J, Thomas, Andrew, Best, Nicky & Spiegelhalter, David. 2000. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, **10**, 325–337.

Mardia, Kanti V., Goodall, Colin, Redfern, Edwin J. & Alonso, Francisco J. 1998. The kriged kalman filter (with discussion). *test*, **7**, 217–282.

Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. 2013. Bayesian computing with inla: New features. *Computational Statistics and Data Analysis*, **67**, 68–83.

Mathéron, G. 1963. Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.

Lütkepohl, Helmut, 2005. *New Introduction to Multiple Time Series Analysis*. Springer.

Montero, J-M, Avilés, G. Fernández & Mateu, J., 2015. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. John Wiley & Sons, first edition.

Morrison, Kathryn T, Shaddick, Gavin, Henderson, Sarah B & Buckeridge, David L. 2016. A latent process model for forecasting multiple time series in environmental public health surveillance. *Statistics in medicine*, **35**, 3085–3100.

Oliver, MA & Webster, R. 1990. Kriging: a method of interpolation for geographical

information systems. *International Journal of Geographical Information Science*, **4**, 313–332.

Onursal, B. & Gautan, S. Vehicular air pollution. experiences from seven latin american urban centers. Technical Paper 373, World Bank, Washington, D.C., 1997.

openair web page. openair: open source tools for air quality data analysis, 2021. URL https://davidcarslaw.github.io/openair/.

Parliament UK. House of commons environment, food and rural affairs, environmental audit, health and social care, and transport committees . improving air quality, 2018. URL https://publications.parliament.uk/pa/cm201719/cmselect/cmenvfru/433/433.pdf.

Polivka, Barbara J. 2018. The great london smog of 1952. *AJN The American Journal of Nursing*, **118**, 57–61.

Pope, C.A. & Dockery, D.W. 1992. Acute health effects of PM10 pollution on symptomatic and asymptomatic children. *Am. Rev. Respir. Dis*, **145**, 1123–1128.

Pope, C.A., Dockery, D.W. & Schwartz, J. 1995. Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology*, **7**, 1–18.

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

Richter, D. & Williams, W. Assessment and management of urban air quality in europe. eea monograph 5. Technical report, European Environmental Agency (EEA), 1998.

Robert, C. P. & Casella, G., 2004. *Monte Carlo Statistical Methods*. Springer, New York, second edition.

Royale College of Physicians. Every breath we take: The lifelong impact of air pollution, 2016. URL https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution.

Rue, H. & Held, L., 2005. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

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*, **71**, 1–35.

Sahu, S.K., Gelfand, A.E. & Holland, D.M. 2006. Spatio-temporal modeling of fine particulate matter. *Journal of agricultural, biological, and environmental statistics*, **11**, 61–86.

Sahu, S.K., Gelfand, A.E. & Holland, D.M. 2007. High-resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association*, **102**, 1221–1234.

Sahu, Sujit K. Bayesian spatio-temporal modelling to deliver more accurate and instantaneous air pollution forecasts. In *UK Success Stories in Industrial Mathematics*, pages 67–74. Springer, 2016.

Shaddick, G, Thomas, ML, Mudu, P, Ruggeri, G & Gumy, S. 2020a. Half the world's population are exposed to increasing air pollution. *NPJ Climate and Atmospheric Science*, **3**, 1–5.

Shaddick, G. & Zidek, J. V. 2014. A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spatial Statistics*, **9**, 51–65.

Shaddick, Gavin, Salter, James M, Peuch, Vincent-Henri, Ruggeri, Giulia, Thomas, Matthew L, Mudu, Pierpaolo, Tarasova, Oksana, Baklanov, Alexander & Gumy, Sophie. 2020b. Global air quality: an inter-disciplinary approach to exposure assessment for burden of disease analyses. *Atmosphere*, **12**, 48.

Shaddick, Gavin, Thomas, Matthew L, Amini, Heresh, Broday, David, Cohen, Aaron, Frostad, Joseph, Green, Amelia, Gumy, Sophie, Liu, Yang, Martin, Randall V *et al.* 2018a. Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology*, **52**, 9069–9078.

Shaddick, Gavin, Thomas, Matthew L, Green, Amelia, Brauer, Michael, van Donkelaar, Aaron, Burnett, Rick, Chang, Howard H, Cohen, Aaron, Dingenen, Rita Van, Dora, Carlos *et al.* 2018b. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**, 231–253.

Shaddick, Gavin & Wakefield, Jon. 2002. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **51**.

Shaddick, Gavin, Yan, Haojie & Vienneau, Danielle. 2013. A bayesian hierarchical

model for assessing the impact of human activity on nitrogen dioxide concentrations in europe. *Environmental and ecological statistics*, **20**, 553–570.

Shaddick, Gavin & Zidek, James V, 2015. *Spatio-temporal methods in environmental epidemiology*. CRC Press.

Shumway, R. H. & Stoffer, D. S. 1982. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*.

Snyder, L.P. 1994. The Death-Dealing Smog over Donora, Pennsylvania": Industrial Air Pollution, Public Health Policy, and the Politics of Expertise, 1948-1949. *Environmental History Review*, **18**, 117–139.

Stein, M.L., 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Verlag: New York.

Tiao, G. C., Box, G. E. P. & Hamming, W. J. 1975. A statistical analysis of the los angeles ambient carbon monoxide data 1955–1972. *Journal of the Air Pollution Control Association*, **25**.

Turnock, S. T., Butt, E. W., Richardson, T. B., Mann, G. W., Reddington, C. L., Forster, P. M., Haywood, J., Crippa, M., Janssens-Maenhout, G., Johnson, C. E., Bellouin, N., Carslaw, K. S. & Spracklen, D. V. 2016. The impact of European legislative and technology measures to reduce air pollutants on air quality, human health and climate. *Environmental Research Letters*, **11**, 024010.

Vicedo-Cabrera, Ana M, Biggeri, Annibale, Grisotto, Laura, Barbone, Fabio & Catelan, Dolores. 2013. A bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in italy. *Geospatial Health*, **8**, 87–95.

Weltgesundheitsorganisation, Organization, World Health & for Environment, European Centre, 2021. *WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.

West, Mike & Harrison, Jeff, 1997. *Bayesian Forecasting and Dynamic Models*. Springer, second edition.

WHO, 2018. *Pollution Management and Environmental Health Program*. https://www.who.int/health-topics/air-pollution.

Wikle, Christopher K. & Cressie, Noel. 1999. A dimension-reduced approach to space-time kalman filtering. *Biometrika*, **86**, 815–829.

Wikle, Christopher K & Hooten, Mevin B. 2010. A general science-based framework for dynamical spatio-temporal models. *Test*, **19**, 417–451.

Wikle, Christopher K., Milliff, Ralph F., Nychka, Doug & Berliner, L. Mark. 2001. Spatiotemporal hierarchical bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, **96**, 382–397.

Wikle, C.K., Berliner, L.M. & Cressie, N. 1998. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117–154.

World Bank, 2016. *Pollution Management and Environmental Health Program.* https://www.worldbank.org/en/programs/pollution-management-and-environmental-health-program.

World Health Organization, 2005. *Air Quality Guidelines: Global Update 2005: Particulate matter, ozone, nitrogen dioxide, and sulfur dioxide.* World Health Organization.

World Health Organization. Ambient air pollution: A global assessment of exposure and burden of disease, 2016. URL https://www.who.int/phe/publications/air-pollution-global-assessment/en/.

Xu, Ke & Wikle, Christopher K. 2007. Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference*, **137**, 567–588.

Zammit-Mangion, Andrew & Cressie, Noel. Frk: An r package for spatial and spatio-temporal prediction with large datasets, 2018.

Zhang, Yang, Baklanov, Alexander, Appel, K Wyat, Ching, Jason, Freitas, Edmilson, Garcia-Pando, Carlos Perez, Henze, Daven K, Jorba, Oriol, Keller, Christoph A, Knievel, Jason C *et al.* 2020. Training materials and best practices for chemical weather/air quality forecasting. *World Meteorological Organization (WMO)*.

Zidek, James V, Le, Nhu D & Liu, Zhong. 2012. Combining data and simulated data for space–time fields: application to ozone. *Environmental and ecological statistics*, **19**, 37–56.

Zidek, Jim, ans Nhu Le, Li Sun & Ozkaynak, Haluk. 2002. Contending with space–time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics*, **13**, 595–613.