Spatio-Temporal Modelling of Air Pollution in Malta

Imran Sheikh

Master's thesis
2018:E47

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

I dedicate this thesis to

my mother *Marianne*

&

my siblings *Rashida*, *Shayan* and *Mahira*

Acknowledgements

First and foremost, I would like to thank the staff at the Environment and Resource Authority (ERA) and the National Statistics Office (NSO) in Malta, for providing me with the air pollution monitoring data and population density data, respectively.

I would like to express my deepest appreciation to my supervisor, Johan Lindström for his constructive advice and feedback on my thesis. It was a pleasure working with him, and without his guidance and encouragement, this thesis would have not been possible.

My sincere gratitude goes to the teaching staff of the Faculty of Science at the University of Lund. They motivated me with the statistical knowledge throughout my Master's programme, to which I am deeply grateful for.

Last but not least, I would like to thank my family and my friends for their continuous ongoing support throughout my Master's programme.

Abstract

Air pollution has become a major worldwide concern due to the high levels of air pollutants emitted from industrial and traffic related activities. Exposure to air pollution has been linked to various negative health effects, ranging from asthma to chronic illnesses. Consequently, analyzing air pollution data has become an essential tool for giving insight about the potential health effects. Spatial statistics is a field where such analysis is possible by dealing with geo-referenced data, i.e., including information about space and time.

This dissertation focuses on spatio-temporal patterns of air pollution in Malta. The main objective is to interpolate concentrations of the nitrogen dioxide (NO_2) pollutant across the country. Two models are presented: a standard Kriging model and a complex spatial-temporal model. The first model uses a Universal Kriging (UK) structure to interpolate concentrations at unobserved locations and/or times. The second model consists of a mean field that incorporates dependence on geographic covariates together with seasonal and long-term trends; and a residual field having a spatial correlation structure.

The models are applied to a dataset consisting of monthly NO_2 concentrations measured at 99 monitoring sites across Malta in 2014-2016. Geographic covariates such as elevation, population density, and distances to coast, roads and industrial areas are used to explain spatial and temporal variations in the NO_2 concentrations. The cross-validated R^2 of the UK and spatio-temporal models are 0.52 and 0.55 respectively. Reconstructions of NO_2 across Malta reveal interesting seasonal and spatial patterns in air pollution. The models are implemented using the statistical software R.

Table of contents

List of figures

List of tables

List of abbreviations

The following is a list of the most used abbreviations throughout the dissertation.

- UK - Universal Kriging
- ML- Maximum Likelihood
- LUR - Land Use Regression
- CV - Cross-Validation
- LTA - Long-Term Average
- RMSE - Root Mean Square Error
- NO_2 - Nitrogen dioxide

List of notation

The following is a list of the most used notation throughout this dissertation. Other notation will be clearly defined.

- $y(s)$ - LTA NO_2 concentrations for location s
- $y(s,t)$ - NO_2 concentrations for location s and time t
- $\mu(s,t)$ - mean field part of $y(s,t)$
- $\varepsilon(s,t)$ - space-time residual field part of $y(s,t)$
- $f_i(t)$ - i^{th} temporal basis functions
- $\beta_i(s)$ - Spatially varying regression coefficients for f_i
- X_i - LUR basis functions for the spatially varying regression coefficients in $\beta_i(s)$
- α_i - Regression coefficients for the i^{th} LUR-basis.
- θ_B - parameters of the covariance structure of β
- θ_ε - parameters of the covariance structure of $\varepsilon(s,t)$
- $\sum_{\beta_i}$ - Covariance matrix describing additional spatial dependence not captured by the geographical covariates.
- $\sum_\varepsilon$ - Covariance matrix for the residuals.
- R^2 - Coefficient of determination

Chapter 1

Introduction

This chapter starts with a brief introduction to air pollution, with focus on the Maltese scenario. Subsequently, a brief introduction to spatial statistics is given, together with the major historical developments of spatial analysis in the field of air pollution monitoring. The problem of interest will be outlined afterwards together with the main objectives of this dissertation. Finally, the structure of the dissertation is briefly explained.

1.1 Air pollution

According to the United Nations (1997), air pollution is defined as:

> *"the presence of contaminant or pollutant substances in the air that do not disperse properly and that interfere with human health or welfare, or produce other harmful environmental effects"*

Air pollution is a major worldwide environmental problem for both ambient (outdoor) and household (indoor) sources. According to the 2016 environmental report from the World Health Organization (WHO), air pollution is the biggest environmental risk to health, accounting for about one in every nine deaths annually. Ambient air pollution alone is responsible for the death of around 3 million people each year. Exposure to air pollutants can lead to a wide range of adverse health outcomes such as cardiovascular disease, respiratory disease, fibrosis, and lung cancer.

There are many sources of ambient air pollution. One of the main sources of outdoor air pollution is fuel combustion from motor vehicles. Industrial facilities such as oil refineries and manufacturing factories also contribute to the growth in air pollution. Other sources of outdoor pollution include heat and power generation and agricultural waste sites.

The air pollutant studied in this dissertation is nitrogen dioxide (NO_2). The gaseous pollutant is known to be a strong respiratory irritant and an important precursor to another toxic pollutant, namely ozone (Chisulio et al. 2011). NO_2 has adverse affects on health, since high concentrations of this air pollutant cause inflammation of the airways and reduced lung function (MITA 2009). Furthermore, high levels of NO_2 are correlated with high levels of mortality (Chen et al. 2012). NO_2 is mainly caused by fossil fuel combustion, particularly in the energy and transport sectors.

1.1.1 Case study – Malta

The country studied in this dissertation, Malta, is an archipelago located in the centre of the Mediterranean Sea in Europe. The archipelago consists of three main islands (Malta, Gozo and Comino), with a population of over 410,000 people (NSO 2011).

Air pollution is a major environmental concern in Malta. According to a 2017 study by the European Environment Agency (EEA), air pollution in Malta is among the worst in Europe, with the archipelago having the fourth highest levels of particles in the air compared to all Member States in Europe. The major source of pollution in the Maltese islands is the increasing use of private cars. According to a 2016 report published by the National Statistics Office (NSO), the country now has over 350,000 licensed motor vehicles. As a result, air pollution has increased particularly by volatile organic compounds, particulate matter and nitrogen oxides.

As a consequence of the growing concern of air pollution worldwide and in Malta, data on air quality is becoming increasingly available. Various researchers throughout the past years have employed methods to analyze air pollution monitoring data. A popular field where statistical methods are employed to analyze such data is spatial statistics, or spatial analysis.

1.2 Spatial statistics

Spatial statistics is a collection of statistical methods that make use of distances and locations to explain the spatial variability of some response variable when making inferences about the process involving the response variable. The location can vary from a street address of an individual to the location of a tree in a forest. Methods involved in spatial statistics include regression, generalized linear models and stochastic process models.

In general, there are three types of spatial data involved in spatial analysis. Cressie (1993) gives the following three main categories:

1. **Geostatistical data** - In this type of spatial data, there are observations of a continuously varying quantity, which are taken at some fixed locations. Examples of this type include measures of temperature at fixed monitoring sites, or air pollution measures at fixed monitoring stations. This type of spatial data will be used in this dissertation.

2. **Lattice data** – This type of data consists of observations indexed over a lattice of points, with the lattice being a regularly spaced or an irregularly spaced region. Examples of this type include images (regularly spaced lattice), or counties in a particular state (irregularly spaced region). Some of the covariates used to explain air pollution in this dissertation will be lattice data.

3. **Point Pattern data** – This type of spatial data consists of observations pertaining to a set of locations where the locations themselves are of interest. In this case, the locations are considered to form a random process. An example of point pattern data would be the location of pine saplings in a Swedish forest.

One of the major developments in the field of spatial statistics is associated with the increase of computerized systems. Geographic information systems (GIS) takes the locations of the study subject, and transform them into geographic coordinates. This technique is known as geocoding. The availability of geocoding and other software has made spatial analysis more accessible to many researchers. Further developments include the use of Bayesian methods in spatial analysis due to advances in Markov Chain Monte Carlo (MCMC) algorithms.

1.2.1 Development of spatial statistics in air pollution monitoring

Since the 1990's, there has been an enormous growth in the statistical models and techniques to analyze spatial data in the field of air pollution monitoring. One of the first researchers was Guttorp et al. (1994), who evaluated ozone data collected in connection with a model study of ozone transport in California. Carroll et al. (1997) formulated a spatial-temporal model for hourly ozone levels in order to predict ozone levels at several locations in Harris County between 1980 and 1993. Haas (1995) suggested a prediction method to evaluate a non-stationary spatio-temporal process. The method was applied to observations on seasonal, rainfall-deposited sulfate in the United States during a six year period.

In recent years, hierarchical Bayesian modelling for spatial prediction of air pollution have been developed. Cressie et al. (1999) compared Kriging and Markov-random field models in the prediction of particulate matter concentrations around Pittsburgh. Kibria et al. (2002) predicted particulate matter concentrations in Philadelphia by suggesting a multivariate Bayesian spatial prediction approach. Cocchi et al. (2007) presented a hierarchical Bayesian model for daily average particulate matter concentration levels. Sahu et al. (2011) developed a hierarchical auto-regressive Bayesian model for space-time air pollution data and evaluated the advantages of Bayesian modeling over other modeling methods with a real data example on monitoring ozone pollution.

Various authors have gone into analyzing the spatial behaviour of the nitrogen dioxide (NO_2) pollutant. Madany and Danish (1993) reported seasonal and spatial variations in the ambient air concentration of NO_2 throughout Bahrain in 1992. Monitoring sites were chosen to include urban areas with high traffic density, suburban areas with low traffic density, commercial, and industrial areas. Lindström et al. (2013a) presented a spatio-temporal model to predict long-term average concentrations of nitrogen oxides (NO_x) in the Los Angeles area during a ten-year period. The objective for the paper was to investigate the relationship between chronic exposure to air pollution and cardiovascular disease. Amini et al. (2016) developed annual and seasonal spatial models for ambient oxides of nitrogen, including NO_2 in the city of Tehran, Iran, using 2010 data from 23 fixed monitoring stations.

A number of authors have employed a number of models to analyze air pollution in Malta. Zammit et al. (2011) suggested a spatio-temporal model using NO_2 and benzene data. Camilleri (2013) mapped the concentrations of NO_2 and identified patterns of the pollutant over space. Zammit (2013) developed a number of spatio-temporal models to analyze air pollution patterns associated with traffic, aiming to capture the temporal and spatial relationships between sites in Malta and Gozo.

1.3 Problem of interest

For the purpose of this dissertation, we will be focusing on spatio-temporal models in the area of air pollution monitoring. We will delve into questions which are of crucial importance to the Maltese environmental sector - Is there a significant link between high levels of air pollution and traffic and/or industrial related areas? Do localities with high population density generate a higher concentration of air pollution? Do the levels of air pollution change depending on the season?

The main interest in this dissertation is the interpolation of nitrogen dioxide across Malta. Interpolation of air pollution data is essential since it gives insight about the potential health effects of the pollutant. Furthermore, the government might take measures in order to control the level of the pollutant. In general, we need modelling tools to be able to interpolate air pollution concentrations. The main purpose of this dissertation is to develop such tools so that they can be used in the future to answer the questions above.

The main objectives of this dissertation are to:

- Study spatio-temporal models for air pollution monitoring
- Identify any geographic covariates affecting the pattern and behaviour of NO_2 across Malta
- Analyze spatially and temporally concentrations of NO_2 across Malta

1.4 Dissertation structure

Apart from this introductory chapter, the dissertation contains four additional chapters, each dealing mainly with the theory and application of spatio-temporal models.

Chapter 2 provides an outline of how air pollution is monitored in Malta through the use of monitoring stations and passive diffusion tubes. This chapter also introduces the NO_2 dataset, focusing on the collection and organization of the data for this dissertation.

In Chapter 3, the spatio-temporal framework is presented. First, the standard Kriging model is theoretically outlined, Then, the spatio-temporal model is theoretically described. For each model, parameter estimation, predictions and model assessment are explained.

In Chapter 4, the spatio-temporal models described in the previous chapter are applied to the NO_2 dataset. Corresponding outcomes are presented together with interpretation of results. Reconstructions for the predictions are also provided in this chapter.

Finally, the dissertation ends with Chapter 5, whereby a summary of the most important concepts is discussed together with an outline of the key findings in the dissertation. Furthermore, limitations of the research are provided together with possible suggestions for similar future studies.

Chapter 2

Air pollution in Malta

In this chapter, a description of the air pollution monitoring in Malta is outlined, including monitoring stations and passive diffusion tubes. Then, a description of the dataset is given together with information on how it was collected and organized for this dissertation.

2.1 Air pollution monitoring in Malta

Monitoring of air pollution is undertaken in Malta through two methods - real-time air monitoring stations, and passive diffusion tubes. All these monitoring sites are operated by the Environmental Resource Authority (ERA), a governmental agency in Malta whose aim is to safeguard the environment to achieve a sustainable quality of life.

2.1.1 Monitoring stations

The real-time air monitoring stations determine concentrations of most pollutants every quarter of an hour (ERA 2018). Pollutants monitored by the stations include nitrogen dioxide, carbon monoxide, volatile organic compounds and particulate matter. The pollutant monitored in each station depends on the nature of the station, its location and purpose.

In Malta, there are currently four real-time stationary measuring stations (ERA 2018). One of them is a traffic site in the locality of Msida, close to the capital city Valletta, which records pollution levels determined mostly by the emissions from nearby traffic. There are two urban stations in the localities of Zejtun and Attard (see map in **Appendix A2**). In this type of monitoring station, pollution levels are not influenced significantly by any single source or street, but rather by a combination of many sources. Finally, a rural background site is located on the island of Gozo, outside of urban areas,

2.1.2 Passive diffusion tubes

Apart from monitoring stations, concentrations of air pollutants can also be measured using a passive diffusion tube system (Camilleri 2013). Passive diffusion tubes are simple and cost-effective devices for air quality monitoring. The low operational costs and ease of usage makes these passive samplers ideal for monitoring the concentration of air pollutants across large areas. According to Nash and Leit (2010), these tubes can also measure much lower concentrations of the same pollutants if exposure time is extended. Therefore, passive diffusion tubes can be useful for indoor and outdoor air quality studies where the objective is to identify locations where average concentrations are particularly low or high.

In Malta, the passive diffusion tube network was introduced in 2004 to have a better spatial coverage over Malta. ERA has installed a number of diffusion tubes measuring the concentration of different air pollutants in many localities throughout Malta and Gozo. The tubes are exposed for a period of 4 weeks after which they are sent to a laboratory for analysis.

2.2 NO_2 Dataset

The dataset analyzed in this dissertation consists of 4-week average concentration levels of the nitrogen dioxide measured at 94 passive diffusion tubes and 5[1] fixed monitoring stations located in Malta and Gozo for the years 2014-2016 (Figure 2.1). Apart from the concentration measurements, the dataset includes the latitude and longitude of each monitoring site.

All in all, there are 33 4-week exposure periods for each monitoring network in the dataset - 13 observations in 2014, 10 observations in 2015 and 10 observations in 2016. To encode the timing of each measurement period, we used the middle date of each exposure period, e.g. the 24th of December 2013 was taken to represent the middle date of the first exposure period 9 December 2013 - 6 January 2014.

2.2.1 Geographic covariates

In order to model the air pollutant concentrations, and to predict at unobserved locations, a number of geographic covariates were added to the NO_2 dataset. These include elevation, distance to coast, distance to trunk, primary and secondary roads, distance to industrial areas and population density.

[1] In the dataset, there are five fixed monitoring stations – the other fixed monitoring station located at Kordin has since been shut down.

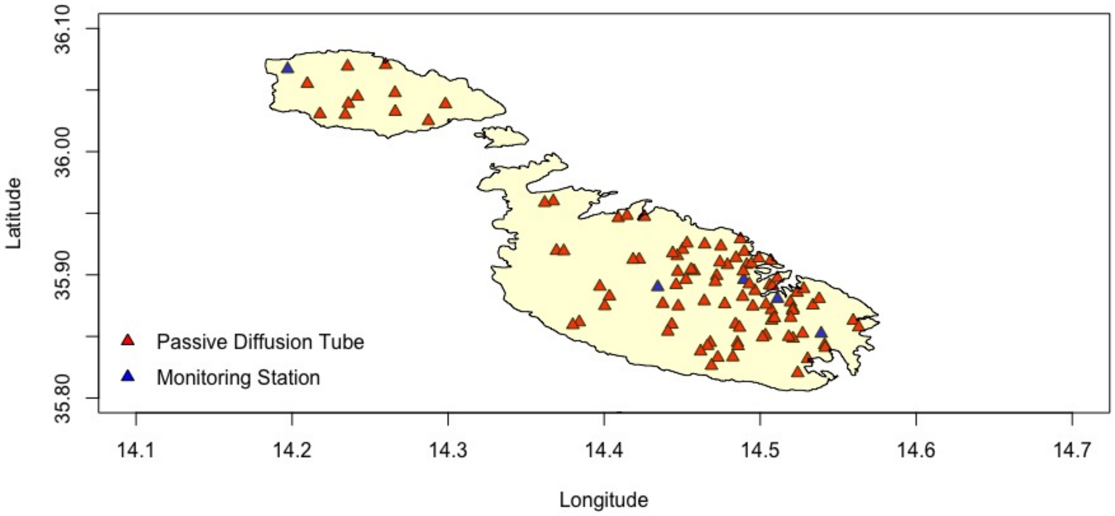


Fig. 2.1 Location of monitoring diffusion tubes and stations in the Maltese Islands. The blue and red triangular dots correspond to the monitoring stations and diffusion tubes respectively.

Spatial information for the elevation was given in the form of Shuttle Radar Topography Mission (SRTM) gridded data from Becker et al. (2009), with average elevation for each gridcell. The elevation in Malta is shown in plot a) from Figure 2.2. As can be seen the elevation is higher in the South Western part of Malta. The elevation for each monitoring site was assigned according to the corresponding gridcell.

The coastline data was provided as a set of hierarchically arranged closed polygons from Wessel and Smith (1996). The distance to coast for each monitoring site was determined by calculating the distances from the site to each polygon data point, and then taking the smallest distance.

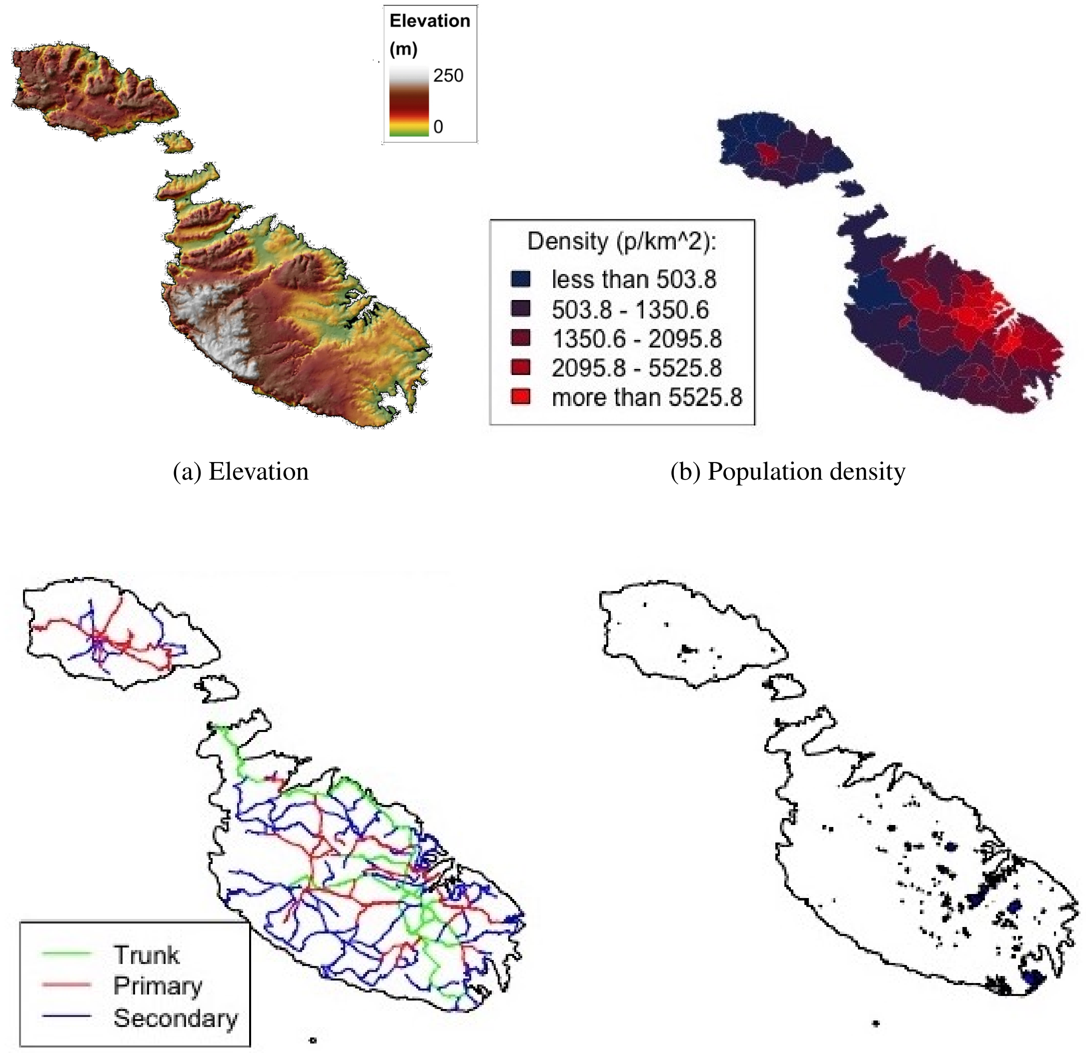
The population density was provided in shapefile format from the National Statistics Office (NSO) in Malta. The shapefile includes 68 localities in Malta and Gozo, and the population density in each locality is given. The population density (measured in km^2) is calculated by dividing the number of people in the locality by the total locality area. According to the 2011 Census of Population Housing report published by the NSO, Malta had a population density of over 1,300 people per square kilometer, making it one of the most densely populated countries in the world. Plot b) from Figure 2.2 shows the population density in Malta and

Gozo. As can be seen, the region including the Southern and Northern harbour districts, have a higher population density (Malta is divided into 6 districts, and each district has a number of localities. Maps of Malta visualizing the 6 districts and the 68 localities are shown in **Appendices A1** and **A2**). Furthermore, Gozo is less densely populated than Malta.

For the NO_2 dataset, each monitoring site was assigned the population density of the locality in which it is situated. So for instance, the passive diffusion site ATD1 and the monitoring station ATD2 are both situated in the locality of Attard, and thus were each assigned the population density of that locality.

Information about major roads (Figure 2.2 (c)) and industrial areas (Figure 2.2 (d)) were provided in shapefile format using the OpenStreetMap online tool[2]. The major roads are divided into three classes namely trunk, primary and secondary roads. The first class refers to high-capacity urban roads which are typically divided. Primary roads are low-to-moderate capacity roads which move traffic from one locality to another. Secondary roads refer to major roads within a locality.

[2] http://download.geofabrik.de/europe/malta.html



(a) Elevation

(b) Population density

(c) Major roads

(d) Industrial areas

Fig. 2.2 Geographic covariates. Plot (a) shows the elevation. Grey areas have the highest elevation above sea level whilst green areas have the lowest elevation. Plot (b) shows the population density. Blue areas have a low population density whilst red areas have a high population density. Plot (c) shows the road network in Malta. Trunk, primary and secondary roads are represented by green, red, and blue lines respectively. Plot (d) shows the industrial areas in Malta.

Chapter 3

Spatio-temporal modelling

In this chapter, we will lay the theoretical foundations of the spatio-temporal framework that will be used to study the NO_2 concentration levels in the Maltese Islands. First, we will describe the standard Kriging model which will be applied to the long-term average NO_2 concentrations. Then, we will shift to the spatio-temporal model which consists of temporal basis functions. For each model, theoretical details about parameter estimation, predictions and prediction accuracy will be outlined.

3.1 Standard Kriging model

In this section, we will theoretically describe a Kriging model used to analyze the long-term average (LTA) concentrations of NO_2 at each monitoring site. The method is named after D. G. Krige, a South African mining engineer who was the first person to construct and apply them (Zimmerman and Stein 2010). Before describing the model, we will introduce some notation.

Let $y(s,t)$ denote logged NO_2 concentrations at location s and time t. Let N and T denote the total number of observations and the number of observation time points, respectively. Let n and p denote the number of observed monitoring sites and the number of geographic covariates, respectively. The LTA $y(s)$ at each location s is given by:

$$y(s) = \frac{\sum_{t=1}^{T} y(s,t)}{T} \tag{3.1}$$

A linear regression model for the observations in (3.1) is constructed as:

$$y(s) = X(s)\beta + \varepsilon(s) \tag{3.2}$$

where $X(s)$ is an $n \times p$ matrix denoting the geographic covariates used in the regression model, and β is a p-dimensional vector of regression coefficients. $\varepsilon(s)$ is an error term which is assumed to be normally distributed with mean 0 and variance τ_ε^2. In matrix notation, (3.2) can be written as:

$$Y = X\beta + E \tag{3.3}$$

where $E \in \mathrm{N}(0, \tau_\varepsilon^2 I)$. We will denote this model as the LTA regression model.

Reasonable estimates for β and τ_ε^2 in (3.2) are

$$\hat{\beta} = (X^T X)^{-1}(X^T Y) \qquad \hat{\tau}_\varepsilon^2 = \frac{||Y - X\hat{\beta}||_2^2}{n-p}$$

If we let y_0 denote an unobserved location and x_0 a covariate observed at y_0, then the conditional expectation of y_0 given $\hat{\beta}$ is:

$$\mathbb{E}(y_0|\hat{\beta}) = x_0\hat{\beta} \tag{3.4}$$

Accounting for the uncertainty in the regression coefficients, the prediction variance of y_0 given τ_ε^2 is:

$$\mathbb{V}(\hat{y}_0|\tau_\varepsilon^2) = \tau_\varepsilon^2 + x_0\mathbb{V}(\hat{\beta}|\tau_\varepsilon^2)x_0^T \tag{3.5}$$

where $\mathbb{V}(\hat{\beta}|\tau_\varepsilon^2) = \tau_\varepsilon^2(X^T X)^{-1}$. The model (3.2) assumes independent and identically distributed (i.i.d.) errors. However, to accommodate for spatial variability in the data, we can shift from a linear regression model to the following Gaussian random model as described in Zimmerman and Stein (2010):

$$y(s) = \mu(s) + \eta(s) + \varepsilon(s) \tag{3.6}$$

where $\mu(s) = X(s)\beta$ is the mean (or trend) component of the model. In this dissertation, we will focus on a Universal Kriging (UK) approach where β is assumed to be unknown. The second term of (3.6), $\eta(s)$ is a zero-mean stochastic process with a parametric stationary covariance function $\mathbb{C}(||s_i - s_j||;\theta)$, where s_i and s_j denote two spatial locations for $i \neq j$, and

θ is a vector of unknown parameters - partial sill (σ^2) and range (ϕ). $\varepsilon(s)$ are uncorrelated measurement errors (or nugget effect) with variance τ_ε^2 (nugget variance). Now, (3.6) gives a multivariate Gaussian model for the observations:

$$Y \in \mathrm{N}(X\beta, \Sigma) \tag{3.7}$$

where Σ is defined as:

$$\Sigma = \Sigma_\eta + \Sigma_\varepsilon = \Big[\mathbb{C}(||s_i - s_j||; \theta)\Big]_{ij} + \tau_\varepsilon^2 \boldsymbol{I}. \tag{3.8}$$

For notation purposes, we can collect the covariance parameters in $\Psi = \{\theta, \tau_\varepsilon^2\}$. We will denote this model the LTA UK model.

3.1.1 Parameter estimation

Parameter estimates for the LTA UK model can be obtained via maximum likelihood (ML) by maximizing the log-likelihood of (3.7). Since Y is Gaussian distributed with mean $X\beta$ and variance Σ, then its probability density function $f(Y)$ is:

$$f(Y) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma(\Psi)|^{\frac{1}{2}}} \exp(-\frac{1}{2}(Y - X\beta)^T \Sigma(\Psi)^{-1}(Y - X\beta)) \tag{3.9}$$

The log-likelihood of (3.7) $l(\Psi, \beta|Y)$, is obtained by taking the logarithm of (3.9). Thus, the likelihood of (3.7) is:

$$l(\Psi, \beta|Y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\Psi)| - \frac{1}{2}(Y - X\beta)^T \Sigma^{-1}(\Psi)(Y - X\beta) \tag{3.10}$$

Parameter estimates then can be obtained by maximizing (3.10) with respect to β and Ψ:

$$\hat{\Psi}, \hat{\beta} = \underset{\Psi, \beta}{\arg\max}\ l(\Psi, \beta|Y)$$

Profile likelihood

According to Lindström et al. (2013a, b), estimating parameters of a large dataset using the ML approach may take considerable computational time. One way to speed up the estimation process is to reduce the number of parameters. The profile likelihood successfully reduces the number of parameters by replacing β with its generalized least square (GLS) estimate

β_{GLS}, and thus the problem is reduced to just estimating Ψ. The GLS estimate of β is:

$$\beta_{GLS}(\Psi) = (X^T\Sigma^{-1}(\Psi)X)^{-1}X^T\Sigma^{-1}(\Psi)Y \tag{3.11}$$

Replacing β with (3.11) in (3.10), parameter estimates are obtained by maximizing $l(\Psi,\hat{\beta}|Y)$ with respect to just Ψ:

$$\hat{\Psi} = \underset{\Psi}{\arg\max}\ l(\Psi,\beta_{GLS}(\Psi)|Y)$$

3.1.2 Predictions

Having obtained parameter estimates, predictions can be computed at unobserved locations. The following setup and computations for the predictions are outlined by Cressie (1993). For the LTA UK model predictions, we can divide the data into known (observed) locations Y_k and unknown (unobserved) locations Y_u. Let μ_k and μ_u denote the expectations of Y_k and Y_u respectively. Moreover, denote Σ_{ku} as the cross-covariance structure between observed and unobserved points, and Σ_{kk} and Σ_{uu} as the covariance structure for the observed and unobserved points respectively. Accounting for the observed and unobserved points, the Gaussian model can be written as:

$$\begin{bmatrix} Y_k \\ Y_u \end{bmatrix} \in N\left(\begin{bmatrix} \mu_k \\ \mu_u \end{bmatrix}, \begin{bmatrix} \Sigma_{kk}\ \Sigma_{ku} \\ \Sigma_{uk}\ \Sigma_{uu} \end{bmatrix}\right) = \left(\begin{bmatrix} X_k\beta \\ X_u\beta \end{bmatrix}, \begin{bmatrix} \Sigma_{kk}\ \Sigma_{ku} \\ \Sigma_{uk}\ \Sigma_{uu} \end{bmatrix}\right)$$

where X_k and X_u denote geographic covariates corresponding to Y_k and Y_u respectively. If y_u denotes an unobserved location, then the conditional expectation of y_u given Y_k and θ is:

$$\mathbb{E}(y_u|Y_k,\theta) = X_u\hat{\beta} + \Sigma_{uk}\Sigma_{kk}^{-1}(Y_k - X_k\hat{\beta}) \tag{3.12}$$

where $\hat{\beta}$ is the GLS estimate given in (3.11). The conditional variance of y_u given Y_k, θ and $\hat{\beta}$ is:

$$\mathbb{V}(y_u|Y_k,\theta,\hat{\beta}) = \Sigma_{uu} - \Sigma_{uk}\Sigma_{kk}^{-1}\Sigma_{ku} \tag{3.13}$$

Adding the uncertainty in the regression parameters, the conditional variance of y_u given Y_k and θ becomes:

$$\begin{aligned}\mathbb{V}(y_u|Y_k,\theta) = {} & \mathbb{V}(y_u|Y_k,\theta,\hat{\beta}) \\ & + (X_u^T - X_k^T\Sigma_{kk}^{-1}\Sigma_{ku})^T(X_k^T\Sigma_{kk}^{-1}X_k)^{-1}(X_u^T - X_k^T\Sigma_{kk}^{-1}\Sigma_{ku})\end{aligned} \tag{3.14}$$

3.1.3 Prediction accuracy

The predictive accuracy of the LTA UK model can be assessed using leave-one-out cross validation. In this method, each location is removed from the data set and the LTA at this location is predicted using the remaining locations.

Given the predictions (3.12) and the prediction variances in (3.13) and (3.14) of the LTA UK model, cross-validated statistics such as the root mean squared error (RMSE) and the coefficient of determination R^2 can be computed. The RMSE is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{n}(y(s) - \hat{y}(s))^2} \tag{3.15}$$

where $\hat{y}(s)$ denotes the predicted $y(s)$ observations. R^2 ranges from 0 to 1, and values closer to 1 correspond to less measurement error. This statistic is computed as:

$$R^2 = \max\left(0, 1 - \frac{\text{RMSE}^2_{LTA}}{\mathbb{V}(y(s))}\right) \tag{3.16}$$

3.2 The spatio-temporal model

Having outlined the standard Kriging model, we will now describe theoretically the spatio-temporal model. The model is based on the notion of smooth temporal basis functions and represents one of the many ways that spatio-temporal dependencies can be modelled. The model has been developed in a series of papers including Szpiro et al. (2010) and Lindström et al. (2013a). The latter implemented the model in an `R` package called `SpatioTemporal` (Lindström et al. 2013b). This package will be used to implement the spatio-temporal model in this dissertation.

We will theoretically describe the model using a similar approach presented in Lindström et al. (2013a, b). For notation purposes, we let $y(s,t)$, N and T be defined earlier in section **3.1**. The spatio-temporal model consists of the equation:

$$y(s,t) = \mu(s,t) + \varepsilon(s,t) \tag{3.17}$$

where $\mu(s,t)$ is the space-time mean field and $\varepsilon(s,t)$ is the space-time residual field. Following the methodology presented in Fuentes et al. (2006), $\mu(s,t)$ can be modelled as follows:

$$\mu(s,t) = \sum_{i=1}^{m} \beta_i(s) f_i(t) \tag{3.18}$$

where $f_i(t)$ is a set of known and fixed smooth temporal basis functions for $i = 1, ..., m$ where m is typically a small number of temporal basis functions (including the intercept). It is assumed that we have an intercept, $f_1(t) = 1$, and that the remaining basis functions $f_2(t), ..., f_m(t)$ have mean zero. The $\beta_i(s)$ are spatially varying regression coefficients for the temporal functions.

The $\beta_i(s)$ in (3.17) are modelled as spatial fields with a UK structure, allowing the temporal structure to vary between locations. The trend in the UK structure is constructed as a linear regression on the geographic covariates. The spatial dependence structure is modelled using a set of covariance matrices $\Sigma_{\beta_i}(\theta_i)$. Thus, the model for $\beta_i(s)$ is:

$$\beta_i(s) \in \mathrm{N}(X_i\alpha_i, \Sigma_{\beta_i}(\theta_i)) \text{ for } i = 1, ..., m \tag{3.19}$$

where X_i are $n \times p_i$ design matrices that can incorporate intercept terms and may include different geographic covariates for the different spatial fields. This component can be denoted as a "land use regression" (LUR) component. Here, p_i denotes the number of LUR-basis functions for the i^{th} temporal basis function, for $i = 1, ..., m$. The α_i are corresponding $p_i \times 1$ matrices of unknown regression coefficients, and $\Sigma_{\beta_i}(\theta_i)$ are $n \times n$ covariance matrices. The β-fields are assumed to be a priori independent of each other.

What remains to specify is a model for the residual space-time field $\varepsilon(s,t)$ in (3.17). Since the temporal basis functions in (3.18) should account for the temporal correlation in the data, $\varepsilon(s,t)$ is assumed to be independent in time, but dependent in space. The residual field is modelled using a Gaussian distribution with zero mean process and a spatial covariance as follows:

$$\varepsilon(s,t) \in \mathrm{N}(0, \Sigma_{\varepsilon}^{t}(\theta_{\varepsilon})) \;\text{ for }\; t = 1, ..., T \tag{3.20}$$

where θ_ε is a multi-dimensional covariance parameter and $\Sigma_\varepsilon^t(\theta_\varepsilon)$ is the spatial covariance block matrix:

$$\Sigma_\varepsilon^t(\theta_\varepsilon) = \begin{bmatrix} \Sigma_\varepsilon^1(\theta_\varepsilon) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_\varepsilon^T(\theta_\varepsilon) \end{bmatrix} \tag{3.21}$$

Here, the size of each covariance matrix, $\Sigma_\varepsilon^t(\theta_\varepsilon)$, is given by the number of observations n_t, at time t. The independence of (3.21) is needed for computational efficiency (see Lindström et al. (2013a) for details). Thus, we have assumed a common family of spatial covariance functions for $\varepsilon(s,t)$ and the various spatial fields embedded in $\mu(s,t)$.

Similar to the LTA UK model, the covariance structures in (3.19) and (3.20) are characterized by a range ϕ, a partial sill σ^2, and a nugget τ^2. In this case, the nugget term in the β-field is assumed to be zero, implying a high correlation between the mean value and the seasonal trend at locations adjacent to each other. The parameters of the spatio-temporal model consist of regression parameters for the geographic covariates:

$$\alpha = (\alpha_1^T, ..., \alpha_m^T)^T \tag{3.22}$$

Spatial covariance parameters for the β_i field:

$$\theta_B = (\theta_1, ..., \theta_m) \tag{3.23}$$

where

$$\theta_i = (\phi_i, \sigma_i^2) \text{ for } i = 1, ..., m$$

and spatial covariance parameters for the spatio-temporal residuals:

$$\theta_\varepsilon = (\phi_\varepsilon, \sigma_\varepsilon^2, \tau_\varepsilon^2)$$

For simplification of notation, we can combine all the covariance parameters into a parameter Ψ:

$$\Psi = (\theta_B, \theta_\varepsilon)$$

Combining (3.17) and (3.18), the spatio-temporal model can be written as:

$$y(s,t) = \sum_{i=1}^{m} \beta_i(s) f_i(t) + \varepsilon(s,t) \tag{3.24}$$

In order to simplify the notation of the model, Szpiro et al. (2010) introduced a sparse $N \times mn$-matrix $F = (f_{st,is'})$ with elements

$$F = (f_{st,is'}) = \begin{cases} f_i(t) & s = s' \\ 0 & \text{otherwise} \end{cases}$$

They also introduced the $N \times 1$ vectors $Y = y(s,t)$ and $E = \varepsilon(s,t)$ by stacking the elements into single vectors, varying first s and then t. Therefore,

$$Y = \begin{bmatrix} y(s_1,1) & y(s_2,1) & ... & y(s_1,2) & y(s_2,2) & ... & y(s_n,T) \end{bmatrix}^T$$

$$E = \begin{bmatrix} \varepsilon(s_1,1) & \varepsilon(s_2,1) & ... & \varepsilon(s_1,2) & \varepsilon(s_2,2) & ... & \varepsilon(s_n,T) \end{bmatrix}^T$$

Components of the β-fields are assembled into block matrices as

$$B = \begin{bmatrix} \beta_1(s) \\ \vdots \\ \beta_m(s) \end{bmatrix}, \ X = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_m \end{bmatrix}, \ \Sigma_B(\theta_B) = \begin{bmatrix} \Sigma_1(\theta_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_m(\theta_m) \end{bmatrix} \tag{3.25}$$

Using the aforementioned matrices, the model in (3.17) can be written in matrix form as:

$$Y = FB + E; \tag{3.26}$$

where

$$B \in \mathrm{N}(X\alpha, \Sigma_B(\theta_B)) \ \text{ and } \ E \in \mathrm{N}(0, \Sigma_\varepsilon(\theta_\varepsilon))$$

Here, $X, \Sigma_B(\theta_B)$ and $\Sigma_\varepsilon(\theta_\varepsilon)$ are block diagonal matrices with diagonal blocks $\{X_i\}_{i=1}^{m}$, $\{\Sigma_{\beta_i}(\theta_i)\}_{i=1}^{m}$ and $\{\Sigma_\varepsilon^T\}_{i=1}^{T}$, respectively. Figure 3.1 illustrates the structure FB decomposed into $FX\alpha$ using two temporal basis functions.

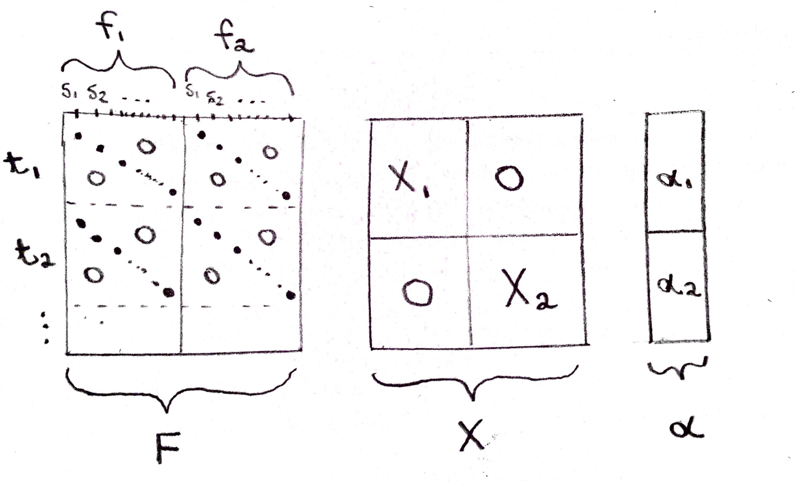


Fig. 3.1 Structure of $FX\alpha$. Note that F will be a large sparse matrix.

It can be noted that (3.26) is a linear combination of independent Gaussians, therefore we introduce the matrices:

$$\tilde{X} = FX \quad \text{and} \quad \tilde{\Sigma}(\Psi) = \Sigma_\varepsilon(\theta_\varepsilon) + F\Sigma_B(\theta_B)F^T \tag{3.27}$$

Consequently. the distribution of Y given the covariance parameters Ψ and the regression parameters α can be written as

$$[Y|\Psi,\alpha] \in \mathrm{N}\left(\tilde{X}\alpha, \tilde{\Sigma}(\Psi)\right) \tag{3.28}$$

3.2.1 Temporal basis functions

The aim of the smooth temporal basis functions, $f_i(t)$, is to capture the temporal variability in the data. These functions are computed using singular value decomposition (SVD) of a data matrix. In this case, $f_1(t) = 1$, thus we end up with $m-1$ smoothed singular vectors. In order to derive the $m-1$ smoothed singular vectors, a $T \times n$ data matrix D is constructed such that:

$$D = \begin{cases} Y & \text{observation } Y \text{ exists} \\ \text{NA} & \text{otherwise} \end{cases} \tag{3.29}$$

To deal with missing observations, an iterative algorithm introduced by Fuentes et al. (2006) is used:

Step 1: Centre and scale each column to mean zero and variance one, and compute the mean of all available observations for each time-point, $\mu_1(t)$. Imputation is done for any missing values in D by fitted values from a linear regression model. In this case, each column of D is regressed onto u_1.

Step 2: Compute the SVD of the new data matrix with any missing values imputed

Step 3: Regress each column of the new data matrix on the first $m-1$ orthogonal basis functions from Step 2. The missing values are then replaced by the fitted values from regression.

Step 4: Repeat from Step 2 until convergence. Convergence is measured by the change in the imputed values between iterations.

The technique of cross-validation (CV) can be used in order to identify the number of smooth temporal basis functions needed to capture the temporal variability in data. In a cross-validation approach, a column of D is held out and smooth temporal functions are computed for the reduced matrix. The functions are evaluated by how well they explain the held out column of D. Repeating for all other columns in D, a set of regression statistics is obtained describing how well the left out columns are explained by smooth temporal functions based on the remaining columns. We can make use of several statistical measures in order to identify a suitable number of temporal basis functions. These include the mean squared errors (MSE), the coefficient of determination R^2, and information criteria such as the Akaike nformation criterion (AIC), and the Bayesian information criterion (BIC).

3.2.2 Parameter estimation

The procedure for parameter estimation is similar to the procedure outlined for the LTA UK model in section **3.1.1**. For the spatio-temporal model, parameter estimates can be obtained via ML by maximizing the log-likelihood of (3.28). Since $Y|\Psi,\alpha$ is Gaussian distributed with mean $\tilde{X}\alpha$ and variance $\tilde{\Sigma}(\Psi)$, then its probability density function $f(Y)$ is:

$$f(Y) = \frac{1}{(2\pi)^{\frac{N}{2}}|\tilde{\Sigma}(\Psi)|^{\frac{1}{2}}}\exp(-\frac{1}{2}(Y-\tilde{X}\alpha)^T\tilde{\Sigma}^{-1}(\Psi)(Y-\tilde{X}\alpha))) \tag{3.30}$$

The log-likelihood $l(\Psi,\alpha|Y)$, is obtained by taking the logarithm of (3.30). Thus, the likelihood of (3.30) is:

$$l(\Psi,\alpha|Y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\tilde{\Sigma}(\Psi)| - \frac{1}{2}(Y-\tilde{X}\alpha)^T\tilde{\Sigma}^{-1}(\Psi)(Y-\tilde{X}\alpha) \tag{3.31}$$

Parameter estimates then can be obtained by maximizing (3.31) with respect to α and Ψ:

$$\hat{\Psi}, \hat{\alpha} = \underset{\Psi,\alpha}{\arg\max}\ l(\Psi, \alpha | Y)$$

As described in section **3.1.1**, the profile likelihood version is obtained by replacing α with its GLS estimate (3.11).

3.2.3 Prediction

Having obtained the estimates of the unknown parameters, the next step is to predict concentrations at unobserved locations and/or times. The different predictions and prediction variances are presented in Lindström et al. (2013b). Before providing predictions for Y, we introduce some notation.

Let X_u and F_u denote the geographic covariates, and temporal basis functions at unobserved locations and/or times, respectively. Furthermore, let B_u denote the collection of β-fields at the unobserved locations. Let $\Sigma_{B,uo}$ and $\Sigma_{\varepsilon,uo}$ denote the cross-covariance matrices between observed and unobserved points, and $\Sigma_{B,uu}$ and $\Sigma_{\varepsilon,uu}$ are the covariance matrices for unobserved points. Using this notation relevant variations on the matrices in (3.27) are

$$\tilde{X}_u = F_u X_u \quad \text{and} \quad \tilde{\Sigma}_{uo}(\Psi) = \Sigma_{\varepsilon,uo}(\theta_\varepsilon) + F_u \Sigma_{B,uo}(\theta_B) F^T \tag{3.32}$$

The model is multivariate normally distributed as seen from (3.28) and the full predictions of unobserved Y's are standard kriging estimates. Therefore, the conditional expectation of the unobserved Y_u given Y and Ψ is:

$$\mathbb{E}(Y_u | Y, \Psi) = \tilde{X}_u \hat{\alpha} + \tilde{\Sigma}_{uo} \tilde{\Sigma}^{-1}(\Psi)(Y - \tilde{X}\hat{\alpha}) \tag{3.33}$$

where $\hat{\alpha}$ is the GLS estimate (3.11) The corresponding conditional variance of Y_u given Y, α and Ψ is:

$$\mathbb{V}(Y_u | Y, \Psi, \alpha) = \tilde{\Sigma}_{uu} - \tilde{\Sigma}_{uo} \tilde{\Sigma}^{-1} \tilde{\Sigma}_{uu} \tag{3.34}$$

Adding the uncertainty in the regression parameters, the conditional variance of Y_u given Y and Ψ is:

$$\begin{aligned}\mathbb{V}(Y_u|Y,\Psi) =&\mathbb{V}(Y_u|Y,\Psi,\alpha)+\\ &(\tilde{X}_u - \tilde{\Sigma}_{uo}\tilde{\Sigma}^{-1}\tilde{X})(\tilde{X}^T\tilde{\Sigma}^{-1}\tilde{X})^{-1}(\tilde{X}_u - \tilde{\Sigma}_{uo}\tilde{\Sigma}^{-1}\tilde{X})^T\end{aligned} \tag{3.35}$$

Given the structure of the model (3.17) with a mean component (3.18) and β-field (3.19), the contribution to any predictions of unobserved Y 's can be decomposed into the following components:

$$\text{Regression component: } \mu(s,t) = \Sigma_{i=1}^{m} X_i \alpha_i f_i(t) \tag{3.36}$$

$$\text{Mean component: } \mu_\beta(s,t) = \Sigma_{i=1} f_i(t)\mathbb{E}(\beta_u|Y,\Psi) \tag{3.37}$$

where

$$\mathbb{E}(B_u|Y,\Psi) = X_u\hat{\alpha} + \Sigma_{B,uo}F^T\tilde{\Sigma}^{-1}(Y - \tilde{X}\hat{\alpha})$$

The components (3.36) and (3.37) play a vital role in model evaluation by highlighting at which level of the model different features of the data are captured. Equations for the predictions of the β-fields and unobserved time points are given in **Appendix A3**.

3.2.4 Prediction accuracy

The predictive accuracy of the model can be assessed through K-fold cross-validation. This method uses part of the available data to fit and a different part to test it. A CV setup is performed by dividing the observed locations into K groups of a reasonably equal size. Figure 3.2 shows a visual diagram of the CV setup when $K = 5$. For the first CV group (Experiment 1), the model is fitted using the other 4 CV groups, and prediction errors are calculated when predicting for the first CV group. A similar approach is done for the other 4 CV groups, and then the 5 estimates of prediction error are combined

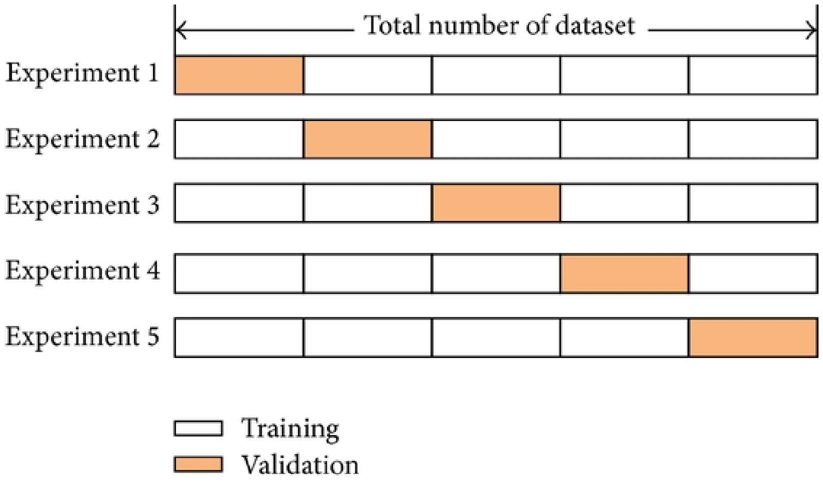


Fig. 3.2 5-fold CV. Source: Zhang et al.(2013)

As outlined in section **3.1.3**, cross-validated statistics such as the RMSE and R^2 can be computed. The only difference is that computations are evaluated for N observations instead of n.

Chapter 4

Analysis

Having described the theory, we will now apply the theory to our NO_2 dataset, described in Chapter 2. We will start off with the spatio-temporal model of the dataset, describing the model and analyzing the results obtained. Finally, we will delve into linear regression and standard Kriging models for long-term average concentrations.

4.1 Initial descriptives

Before analyzing the temporal structure in the data, we evaluated the seasonality and normality in the observations. Simple plotting of the observations at different locations indicated whether there was any seasonal pattern present in the NO_2 observations. Figure 4.1 shows the times series plots of 4 passive diffusion tube locations. Seasonality is evident at all locations, with higher NO_2 concentrations occurring the winter season, and lower NO_2 concentrations occurring during the summer season.

The distribution of the observations was evaluated though a normal Q-Q plot as shown in the two plots in Figure 4.2. The left-hand plot shows the Q-Q plot for the raw observations, whilst the right-hand plot shows the Q-Q plot for the logged observations. In both cases, there were evident heavier tails. In the raw observations, there seemed to be more deviations than the logged observations. Thus, from now onwards, we will focus on logged NO_2 observations for our analysis.

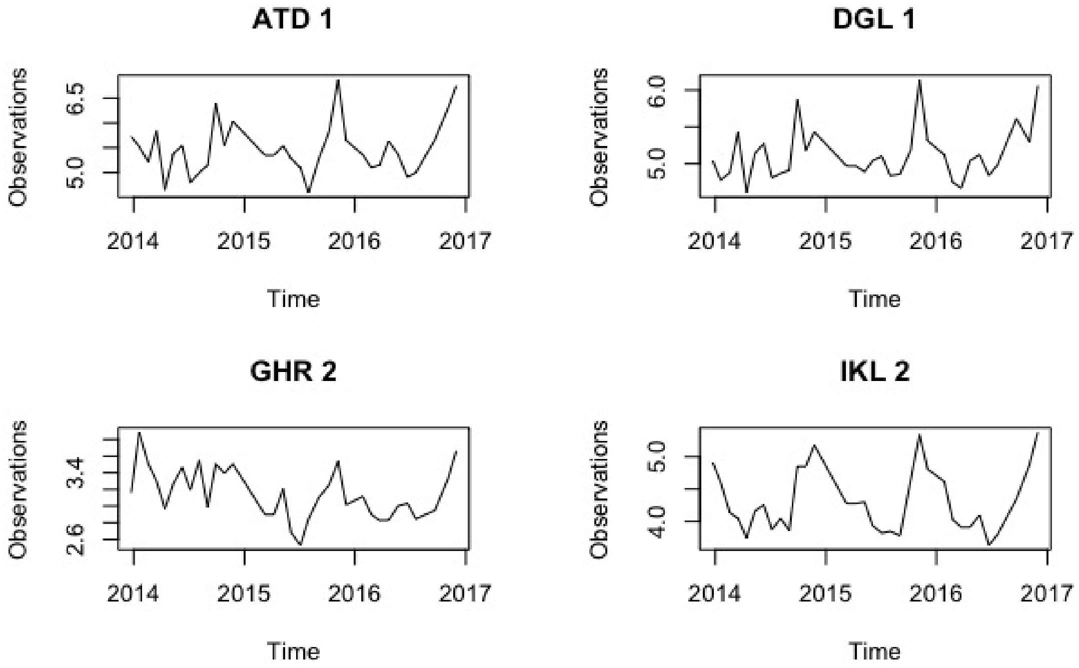
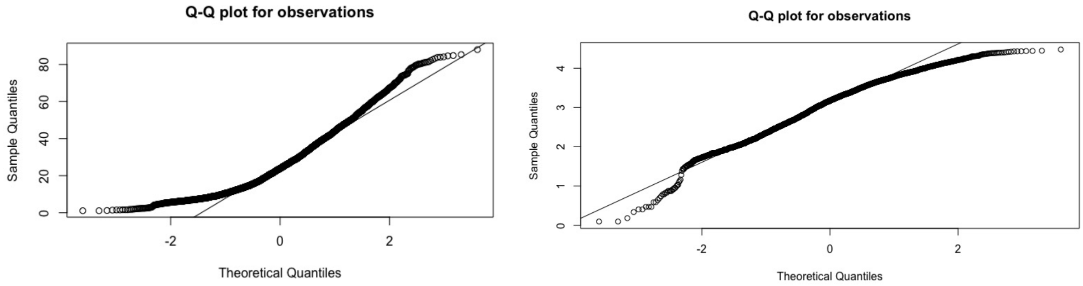


Fig. 4.1 Time Series plots of 4 locations.



(a) Raw data

(b) Logarithmic data

Fig. 4.2 Q-Q plot of observations.

4.2 Temporal basis functions

After analyzing the seasonality and normality assumptions of the observations, the next step was to evaluate the temporal structure and determine the number of spatially varying smooth temporal basis functions. To determine a suitable number of basis functions, the cross-validation technique described in section **3.2.1** was used. Figure 4.3 shows the cross-validation statistics MSE, R^2, AIC, and BIC for each of the number of basis functions evaluated. With regards to information criteria, the basis number having the lowest AIC and BIC should be preferred. In this case, the AIC indicated that the number of basis functions should preferably be 3, whilst the BIC showed that the number of basis functions should preferably be 2. All four statistics seemed to flatten out mostly after 2 basis functions, indicating that 2 basis functions is likely to provide the most efficient description of the temporal variability.

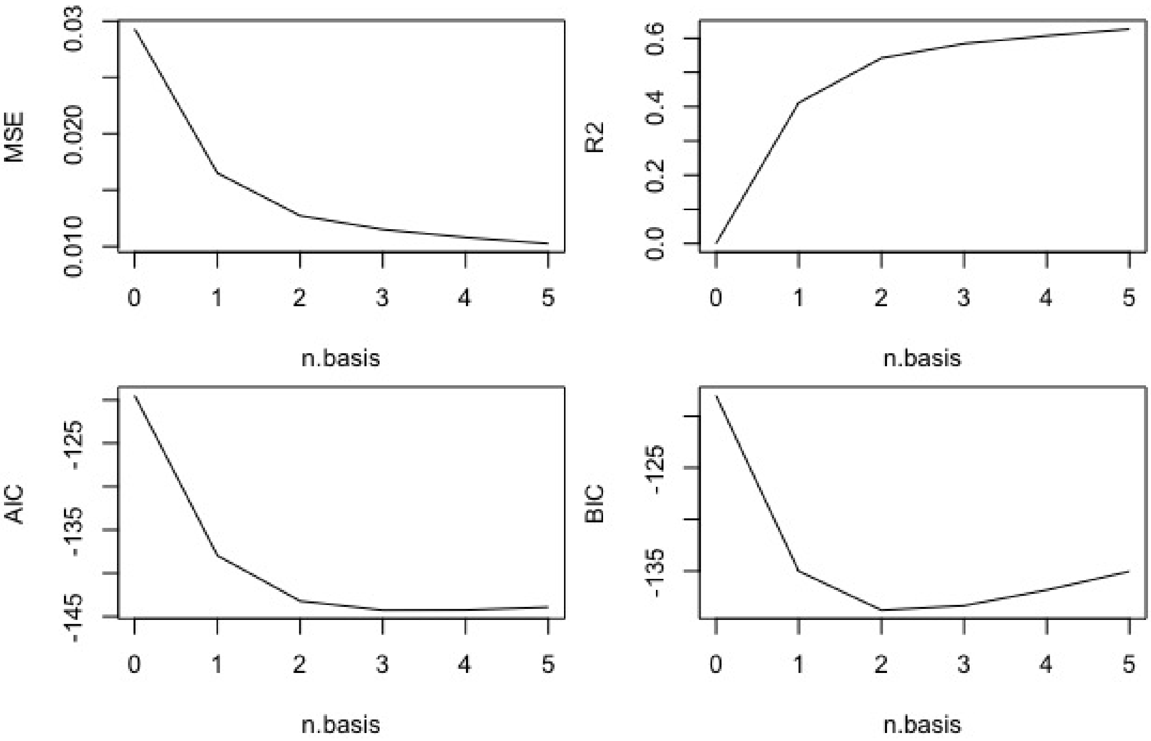


Fig. 4.3 Cross-validated results.

4.3 Modelling the β and ε-fields

After determining suitable temporal basis functions. the type of spatial covariance model to use for each β and ε fields was specified. Furthermore, the LUR covariates for each of β fields were identified.

Since we have 2 temporal basis functions, 2 latent β-fields (β_1 and β_2) and another latent β_0-field corresponding to the intercept were modelled. The three β-fields were estimated by regressing the observations for each site on the temporal trends. The regression coefficients and standard deviations for each of the β-fields were also extracted.

Plotting the regression coefficients against the covariates gave us a rough idea regarding which covariates to include in the β-fields. Figure 4.4 shows two plots. Both plots show the regression estimates of the β_0 field at each location, with 95% confidence intervals, as a function of distance to trunk road. The only difference is that in the first plot, the distance to trunk road is analyzed in its original scale, whilst in the second plot, the same covariate is transformed in logarithmic scale. As can be seen, the second plot shows a clearer negative linear pattern. Thus, the logarithmic version seems to be a more reasonable covariate to include in the β_0-field. Following similar analysis, it was decided to include the log-transformation of the population density, distance to roads and distance to industrial areas in the spatio-temporal model.

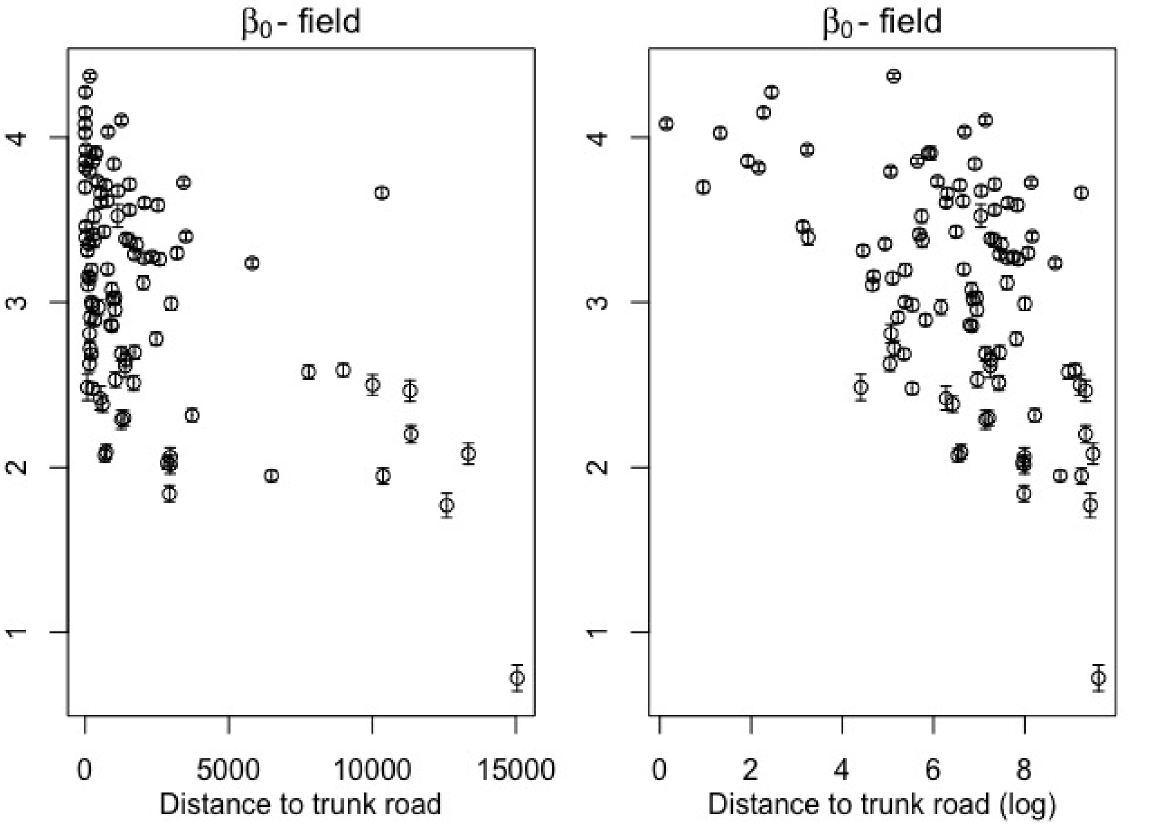


Fig. 4.4 Regression estimates of β_0 field at each location, with 95% confidence intervals, as a function of distance to trunk roads in natural and log scale.

Apart from plotting, a stepwise selection approach based on AIC, and using the `step` function in `R` was used to identify the most suitable covariates to include for each of the β-fields. The second column of Table 4.1 shows the selected variables for the β-fields.

Having obtained covariate specifications for the β-fields, the next step was to determine covariance specifications for both the β and ε-fields. For the latter, we used an exponential covariance with a nugget:

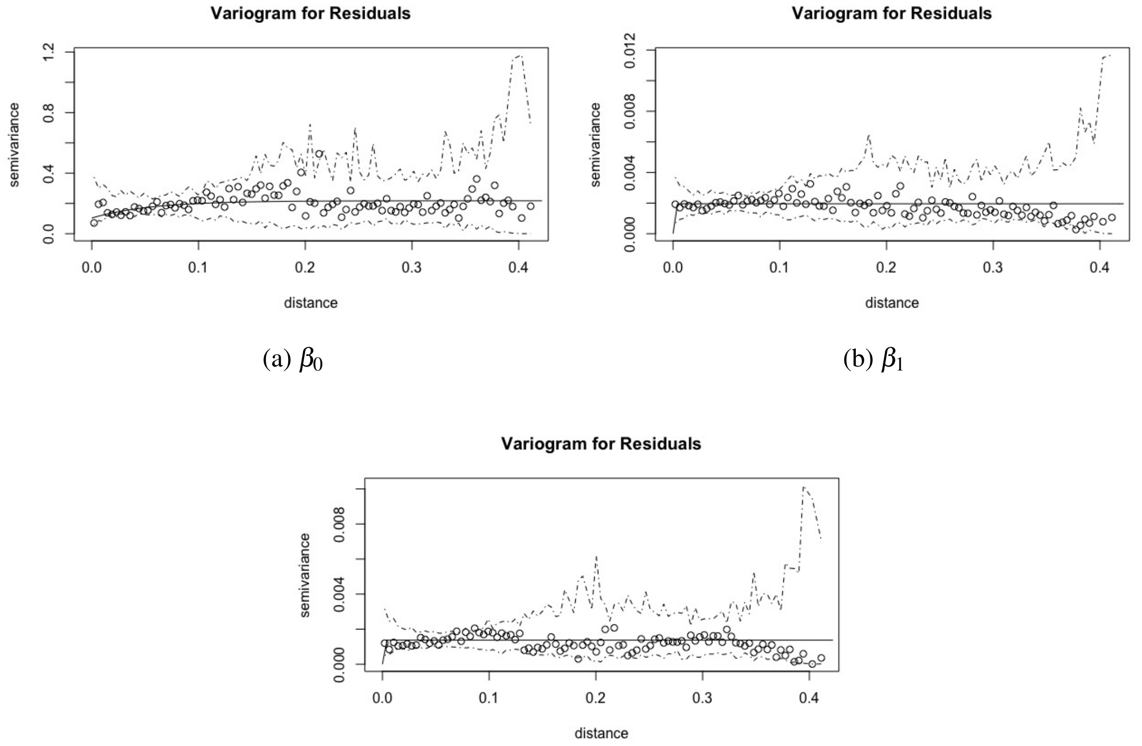
$$\sigma_{\varepsilon}^{2}\exp(-\frac{d}{\phi_{\varepsilon}})+\tau_{\varepsilon}^{2} \tag{4.1}$$

where d denotes the distance between observation locations. Covariance specifications for β_0, β_1, and β_2-fields were determined by analyzing the variogram of the residuals from the regression models of the β-fields, as displayed in Figure 4.5. In all variograms, there seemed to be hardly any spatial dependence, except for a few spatial effect in the β_0-field. Consequently, all three β-fields were modelled using an independent and identically distributed covariance, meaning that only a nugget is included.

Having defined model specifications for the β and ε-fields, we specified a list of coordinates for the observations. In this case, the latitude and longitude of the monitoring sites were taken to be the coordinates used to compute distances between observation locations. Table 4.1 gives a summary of the spatio-temporal model, including covariate selections for the β-fields, and covariance specifications for the β and ε-fields.

4.4 Parameter Estimation

Given model specifications for the β and ε-fields, parameters of the spatio-temporal model were estimated by using the ML approach described in section **3.2.2**. The parameters include the regression coefficients (intercepts and coefficients of the selected covariates) and the 6 covariance parameters shown in Table 4.1. The resulting parameter estimates of the β and ε fields are shown in **Appendix A4**.



(a) β_0

(b) β_1

(c) β_2

Fig. 4.5 Variograms of the β-field residuals. The top and bottom curves shown in each plot are called the variogram envelopes. The envelopes are computed based on permutations of the data values across the spatial locations, meaning that the envelopes are constructed under the assumption of no spatial dependence (Ribeiro and Diggle 2001).

Field	Selected Covariates	Covariance function	Covariance parameters
β_0	Elevation, latitude, distances to trunk, primary and secondary roads	i.i.d.	τ_0^2
β_1	Elevation, population density, distances to trunk, primary and secondary roads, distance to coast and industrial areas	i.i.d.	τ_1^2
β_2	Latitude, longitude, elevation, population density and distance to coast	i.i.d.	τ_2^2
ε		Exponential	$\theta_\varepsilon = (\sigma_\varepsilon^2, \phi_\varepsilon, \tau_\varepsilon^2)$

Table 4.1 Summary of the spatio-temporal model.

4.5 Predictions

Given estimated parameters, predictions were computed for the Gaussian model. These include conditional expectations (3.33) and prediction variances (3.34) of the NO_2 observations. It is assumed that regression parameters are known and prediction variances do not include uncertainties in regression parameters. Contributions to the predictions from the regression component (3.36) and the mean component (3.37) are also computed.

Figure 4.6 shows the predicted and observed data for 4 out of the 99 monitoring sites. It can be noted that in all four locations, the different predictions approximately capture the seasonal variations in the data. The contribution from the regression component seems to capture less seasonality than the other predictions. This can be noted by the deviations from the confidence intervals at certain periods of time.

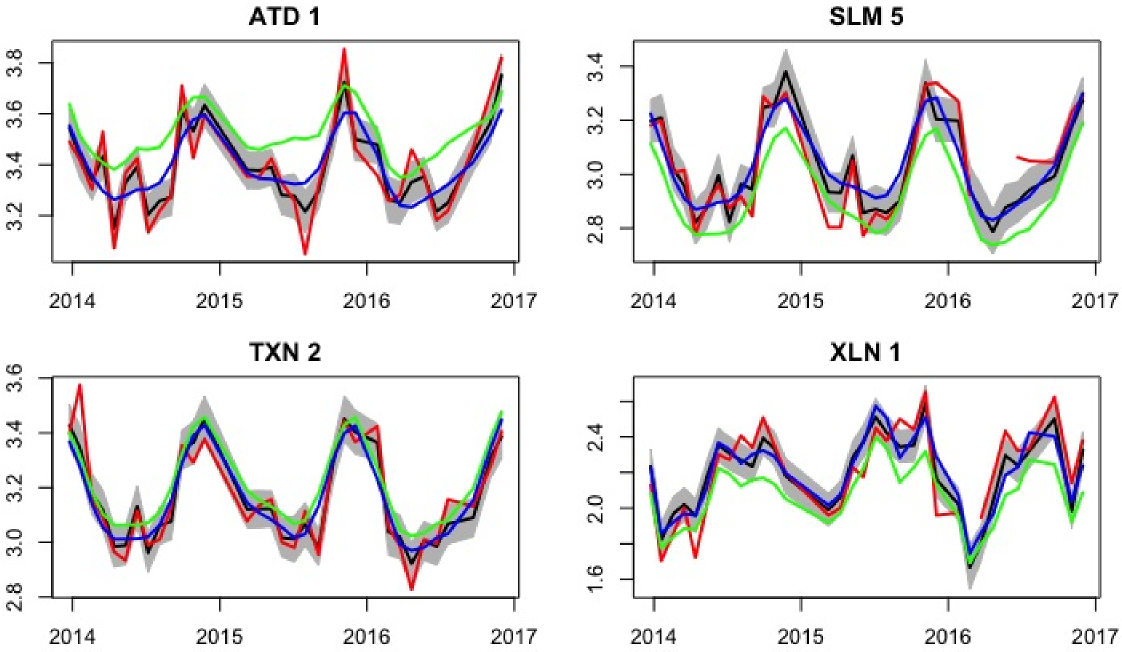


Fig. 4.6 Observations and different predictions in 4 monitoring sites. The red lines denote observations, whilst the black line and gray shading give predictions and 95% confidence intervals at unobserved time-points respectively. The green and blue lines give the contribution to the predictions from the regression and mean components, respectively.

4.6 Cross-validation and Model Assessment

The model's predictive ability was assessed through the cross-validation method described in section **3.2.4**. The observations were split into 6 cross-validation groups. Points closer than 0.01 were forced into the same group. Thus, the number of observations in each group was roughly even. Table 4.2 shows the 6 CV groups and the number of observations in each group.

CV group	1	2	3	4	5	6
Observations	327	677	784	422	473	526

Table 4.2 Cross-Validation groups.

4.6.1 Cross-validated Estimation and Prediction

Having created the CV groups, parameters were estimated for each CV group using the covariance estimates obtained in section **4.4** as starting values. The CV estimation results are shown in Table 4.3. We can see that the parameter estimates for all 6 CV groups converged. The table also gives the optimal log-likelihood value for each estimate.

CV group	Log-likelihood	Convergence
1	4786.74	TRUE
2	4217.15	TRUE
3	3982.43	TRUE
4	4580.00	TRUE
5	4440.47	TRUE
6	4416.75	TRUE

Table 4.3 Cross-validated Estimation results.

Figure 4.7 shows the estimated covariance parameters, on a log-scale, and approximate 95% confidence intervals (red) compared to covariance estimates from the 6-fold cross-validation (box-plots). In this case, there seems to be a reasonable agreement with values and uncertainties for most of the estimates.

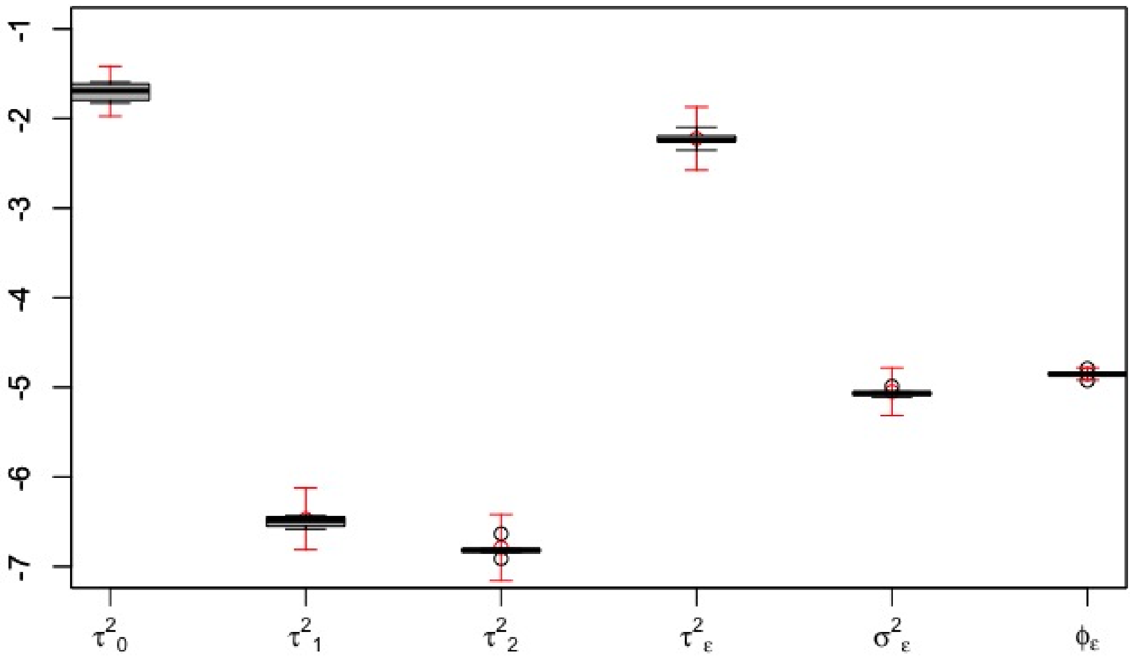


Fig. 4.7 Estimated log-covariance parameters and approximate 95% confidence intervals (red) compared to parameter estimates from 6-fold cross-validation (box-plots).

Given the estimated parameters, predictions were computed for each of the CV-groups. This was achieved by computing the conditional expectations for the left-out observations, given all other observations and the estimated parameters.

4.6.2 Model Assessment

Having obtained CV estimations and predictions, the predictive ability of the spatio-temporal model was evaluated. The statistical measures RMSE and R^2 were used to assess the accuracy of the model.

Table 4.4 shows the CV statistics for all the observations, using three different components of the spatio-temporal model - conditional expectation (3.33), the regression component (3.36) and the mean component (3.37). As can be seen, the R^2 of all components is quite low both indicating that the spatio-temporal model is not of a reasonable good fit and thus, needs to be improved.

Model components	RMSE	R^2
$\mathbb{E}(Y_u \mid Y, \Psi)$	0.46	0.55
$\mu(s,t)$	0.47	0.54
$\mu_\beta(s,t)$	0.47	0.54

Table 4.4 Cross-validation statistics for the observations.

Apart from cross-validation statistics, model assessment was evaluated graphically as shown in Figure 4.8 and Figure 4.9. Plotting all predictions against observations seems reasonable, although some locations exhibit biases. The predicted LTA's match the observations, but with rather large prediction intervals. The width of the intervals is caused by a number of monitoring sites only having a few years of data to average over. Analyzing the residuals for the Gaussian model, (Figure 4.9), we see that they are close to normal, but have slightly heavier tails than expected.

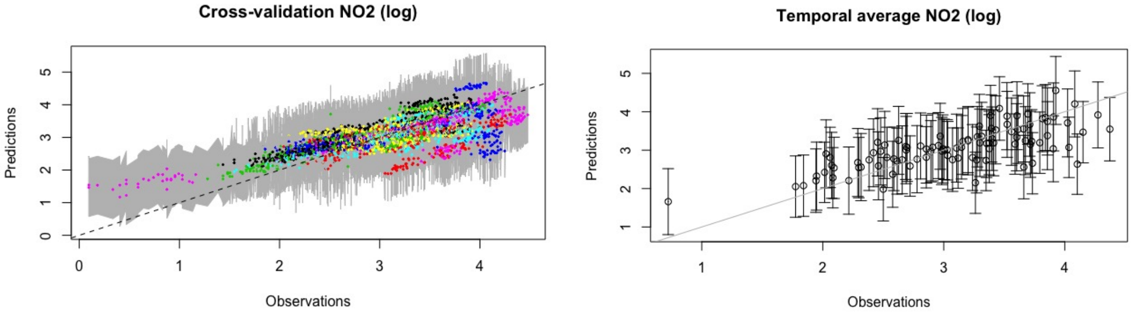


Fig. 4.8 Cross-validation plots. The left-hand side plot compares the CV predictions with the observations. The points are coloured by location and grouping of data from single locations as well as site specific biases can be seen. The right-hand side plot compares the predictions with long-term averages.

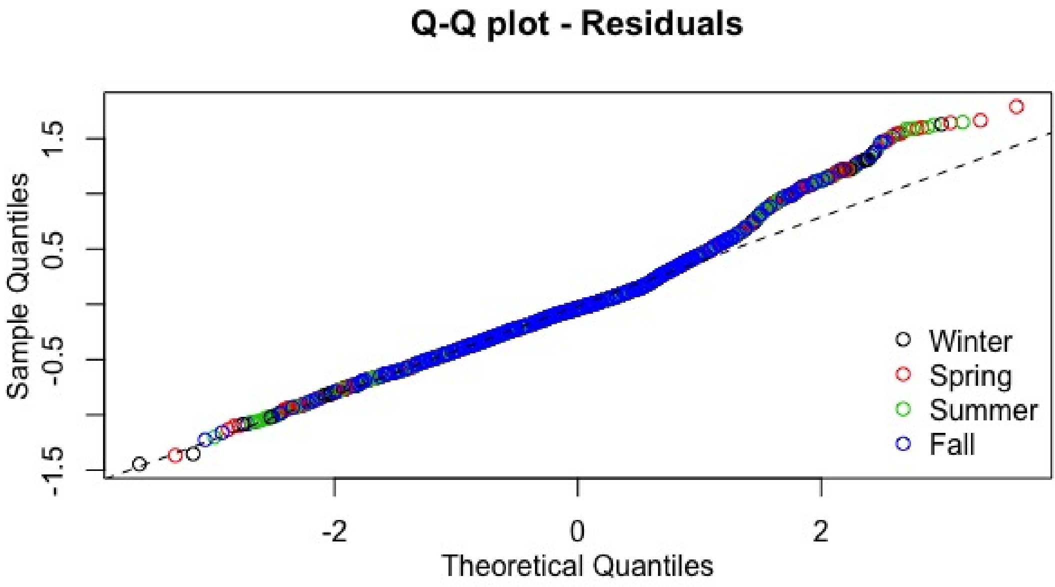


Fig. 4.9 QQ-plot for the residuals. To evaluate seasonal patterns, the plot is colour coded by season - winter (black), spring (red), summer (green) and autumn (blue). The dashed line in the plot gives the theoretical behaviour of normally distributed residuals.

4.7 Reconstruction and prediction uncertainties

In this section, the reconstructed prediction fields are presented. These include the reconstruction fields of the prediction using various components of the spatio-temporal model. The components include the conditional expectations (Figure 4.10), and contributions from the regression (Figure 4.11 plot (a)) and mean (Figure 4.11 plot (b)) part of the model. Furthermore, prediction uncertainties are also plotted (Figure 4.12).

In order to plot the reconstructions, a prediction grid was constructed. A regular latitude-longitude grid with resolution 0.05 covering the whole archipelago of the Maltese islands was created. The LUR covariates for each coordinate were generated using the methods described in **2.2.1**. To avoid issues with the log-transformation due to zeros, distances to coast were truncated from below to the smallest distance to coast from all monitoring sites.

Evaluating the conditional expectation reconstruction field (Figure 4.10), there seems to be higher level of NO_2 concentration levels during the winter season. A major source could be motor vehicles driven by people for work or school purposes, leading to busy traffic roads. It can be seen from all the plots that the most densely populated areas have a higher level of NO_2 concentrations. If we analyze the island of Gozo, we can see that the small island has a lower level of NO_2 concentration than in Malta due to its low population density. Furthermore, Gozo is considered to be greener and more rural with less traffic. Contrary to Malta, the prediction in Gozo seems to be higher in summer than in winter. This could be due to the large numbers of vehicles crossing between Malta and Gozo during the summer period (NSO 2018). A source for lower NO_2 predictions during winter in Gozo could be the clear, fresh weather induced by the north-westerly wind. According to NSO (2011b), this wind is the most common wind direction in Malta, and it is stronger during the winter period.

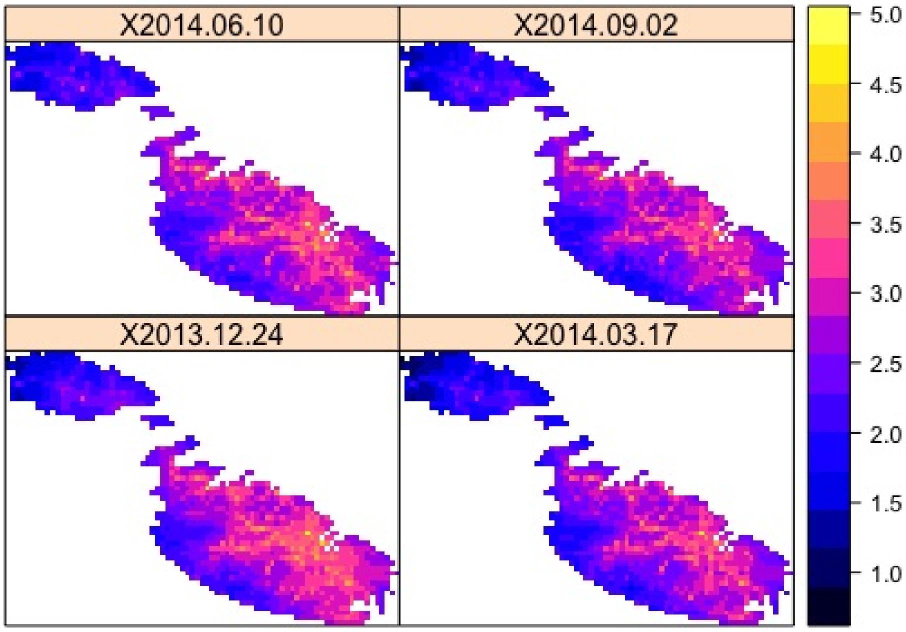
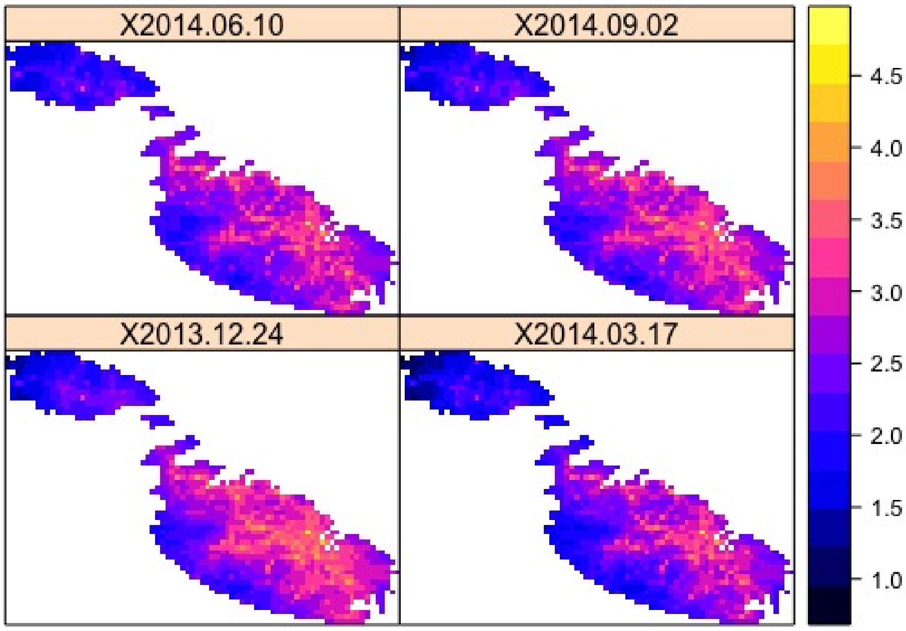
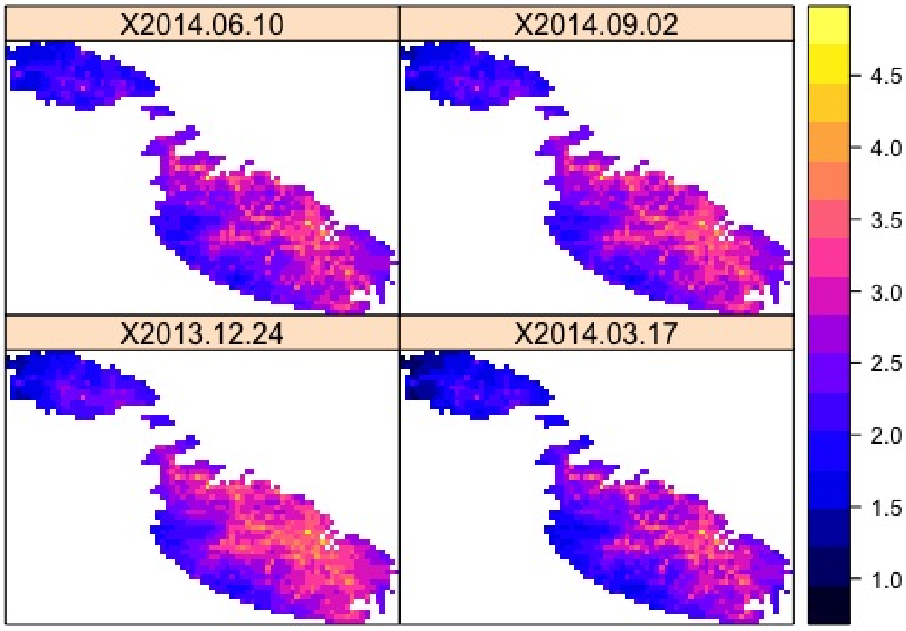


Fig. 4.10 Reconstructed prediction field. In the figure, there are 4 plots with each plot corresponding to a different observation date. Each date corresponds to a different season. The top plots include observations during the summer, whilst the bottom plots include observations from winter. It can be seen from the bottom-left hand plot, that there are areas with high level of predicted NO_2 concentration levels, particularly in the eastern part of the island.

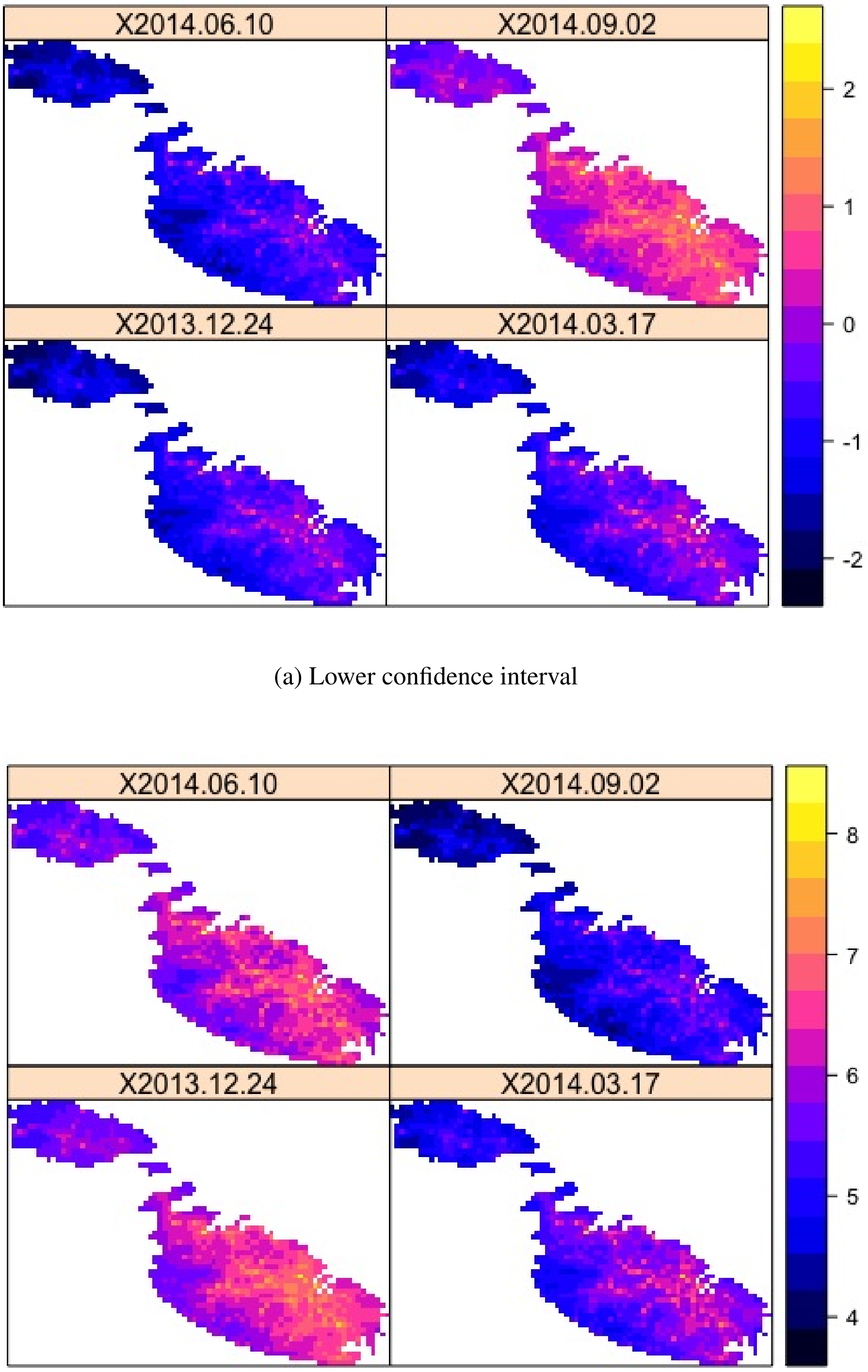


(a) Regression component



(b) Mean component

Fig. 4.11 Reconstructed prediction field using the regression component (a) and the mean component (b) of the spatio-temporal model. All plots look very similar to the reconstructed field in Figure 4.10, indicating that the different components of the spatio-temporal model do not contribute to any spatial effect in the model.



(a) Lower confidence interval

(b) Upper confidence interval

Fig. 4.12 Reconstructed prediction uncertainty (95% confidence interval) plots.

4.8 Long-term averages

After analyzing the spatio-temporal model for the NO_2 observations, the focus is placed on the LTA concentrations at each monitoring site. The LTA regression model and the LTA UK model outlined in section **3.1** will be applied to the long-term averages. This is done to notice any spatial patterns, and to evaluate the differences between this approach and the spatio-temporal model analyzed before.

4.8.1 LTA regression model

The long-term averages were extracted using (3.1). The linear regression model outlined in (3.2) was applied to the LTA observations, where the latter served as the response variable, and the LUR covariates were included as explanatory variables. All covariates were included in the model, and a stepwise approach using AIC was used in order to determine the most suitable covariates to include in the LTA regression model. Table 4.5 shows the selected covariates together with regression estimates.

Figure 4.13 shows the comparison between the true observations (shown in red) and the predicted observations from the LTA regression model (shown in blue). Except for a few minor deviations, the LTA regression model provides a reasonable good fit for the long-term averages.

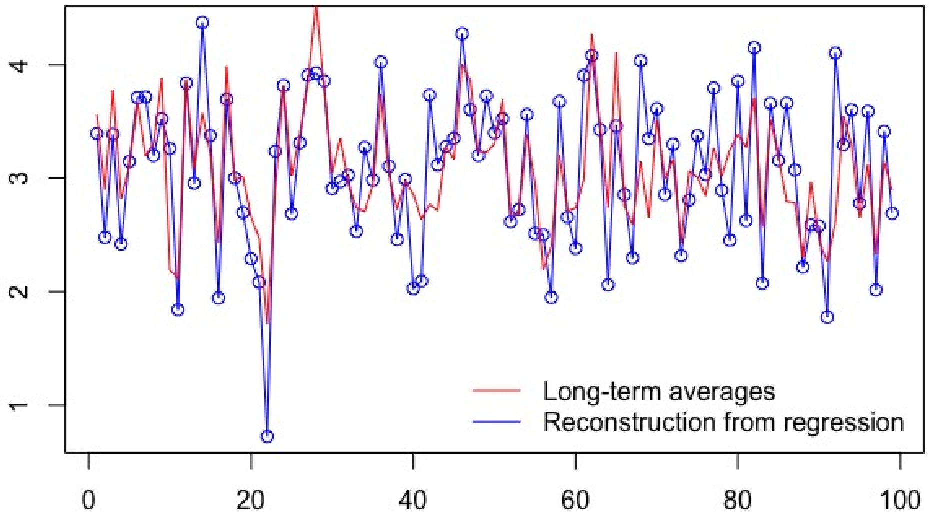


Fig. 4.13 True observations vs LTA regression predictions.

4.8.2 LTA UK model

The variogram of the residuals from the LTA regression model was analyzed in order to detect any spatial dependence. This is shown in Figure 4.14. From the variogram, almost all residual points are enclosed within the variogram envelope, indicating a minimal spatial effect.

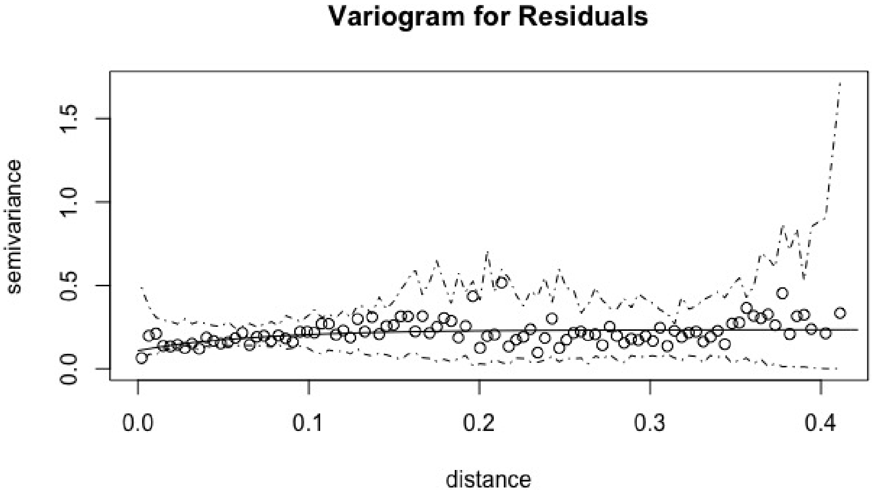


Fig. 4.14 Variogram of LTA regression model residuals.

Despite little spatial structure, the LTA UK model (3.6) was constructed for comparison purposes. As outlined in (3.6), the model consisted of a mean (trend) component $\mu(s)$, a spatial component $\eta(s)$ and an error component $\varepsilon(s)$. In this case, the former consisted of the LUR variables included in the LTA regression model (Table 4.5); $\eta(s)$ consisted of the covariance parameters σ^2 (partial sill) and ϕ (range); and the latter component consisted of τ_ε^2 (nugget variance). Since the exponential function was used to model the ε-field in the spatio-temporal model, the same covariance function was used in the LTA UK model.

Parameter estimation

The regression and covariance parameters for the LTA UK model were estimated using the ML procedure described in section **3.1.1**. The values of the estimated parameters for the 3 components of the LTA UK model are shown in Table 4.5.

Component	Parameters	Selected Covariates	Regression Estimates	UK Estimates
Mean	β_0	(Intercept)	6.047	5.972
	β_1	Elevation	-0.003	-0.003
	β_2	Distance to trunk road	-0.213	-0.212
	β_3	Distance to primary road	-0.151	-0.151
	β_4	Distance to secondary road	-0.116	-0.110
Spatial	σ^2			0.02
	ϕ			0.036
Error	τ_ε^2		0.192	0.183

Table 4.5 Summary of LTA Regression & UK models.

Model assessment

The validity of the LTA UK model was assessed through leave-one-out cross validation as described in section **3.1.3** . Figure 4.15 shows a plot of the LTA observations against the predicted values after performing cross-validation. The predicted values seem to be close to the true observations for most monitoring locations as evidenced by the linear pattern in the plot. However, there are a few points which deviate a bit from each other. A noticeable example is situated at the bottom left hand corner of the plot. This corresponds to station GRB 2 in Gozo, which recorded the lowest NO_2 concentrations throughout the whole 3 years.

The R^2 value for the LTA UK model was computed using (3.16), generating a value of approximately 0.52. This is quite low, indicating that the LTA UK model needs to be improved.

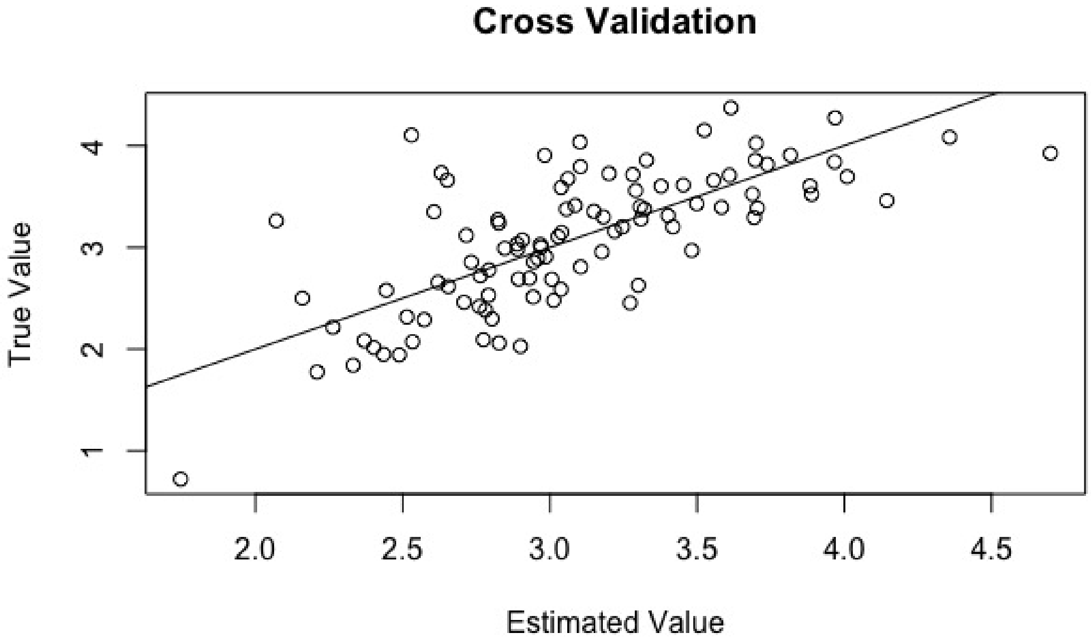
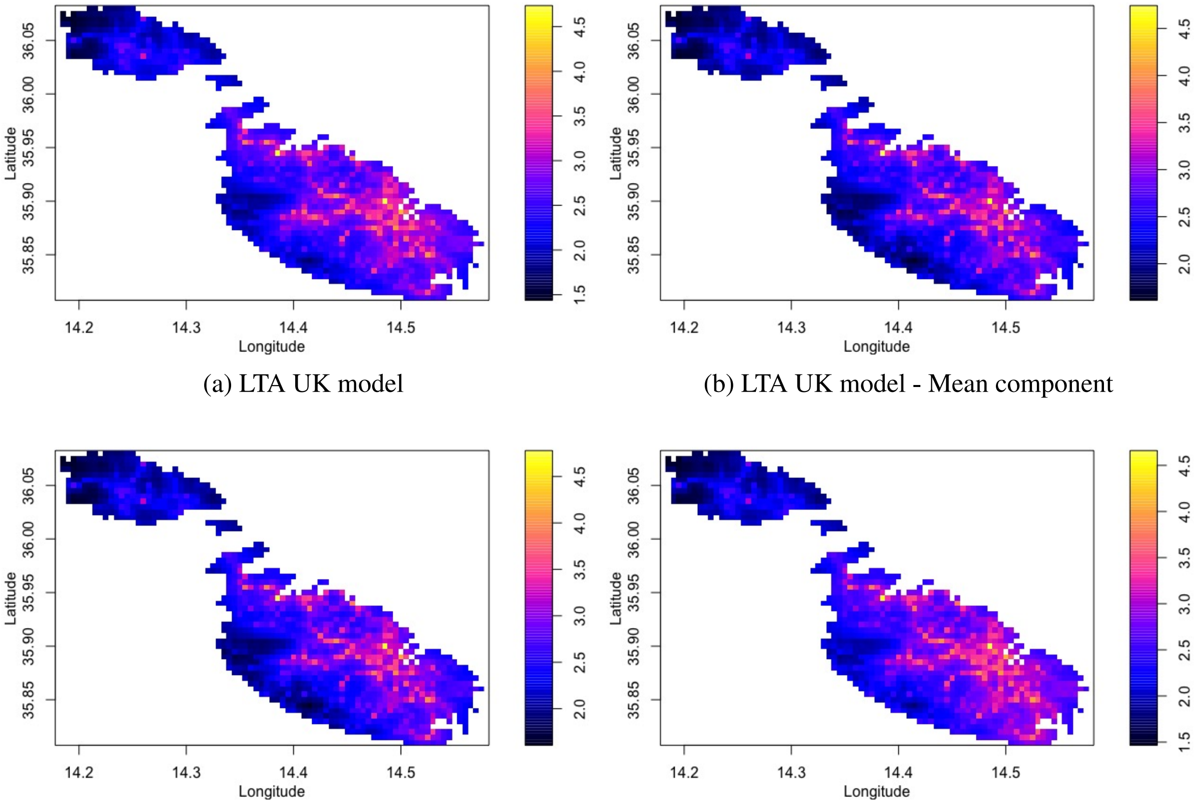


Fig. 4.15 Cross-validation for the LTA UK model.

4.8.3 Predictions and reconstructions

Using the LTA regression model parameters and the LTA UK model parameters from Table 4.5, predictions can now be computed at unobserved target locations over the entire Maltese Islands. For the LTA regression model, conditional expectations and variances are computed as shown in (3.4) and (3.5) respectively. For the LTA UK model, conditional expectations and variances are computed as shown in (3.12) and (3.13), respectively.

In order to perform predictions, the prediction grid described in section **4.5.1** was used. Figure 4.16 shows the reconstruction plots of the predicted values. All plots look very similar to each other, indicating that both the LTA UK model and the LTA spatio-temporal model are capturing minimal additional spatial effect. As can be seen in all of the plots, higher predictions (shown in red and yellow) occur in areas characterized mostly by trunk and primary roads whilst lower predictions (shown in blue and black) occur in areas characterized mostly by secondary roads.



(a) LTA UK model

(b) LTA UK model - Mean component

(c) LTA Regression model

(d) LTA Spatio-temporal model

Fig. 4.16 Reconstructions of the predicted LTA observations. Plot (a) shows the LTA UK model predictions; plot (b) shows the predictions using just the mean component of the LTA UK model; plot (c) shows the predictions of the LTA regression model and plot (d) shows the predicted LTA values extracted from the spatio-temporal model.

Figure 4.17 shows the reconstructed field of the prediction uncertainty (standard deviation) of the LTA regression model. Elevation seems to be a contributing geographic variable in the prediction uncertainty, since the standard deviation is higher for areas with a high elevation. Furthermore, there seems to be points (shown in black) which clearly stand out from the rest. These points have a lower standard deviation compared to other neighbouring prediction points. After closely analyzing the LUR covariates associated with these prediction points, it turns out that such points have much smaller distances to primary roads compared to other prediction points nearby.

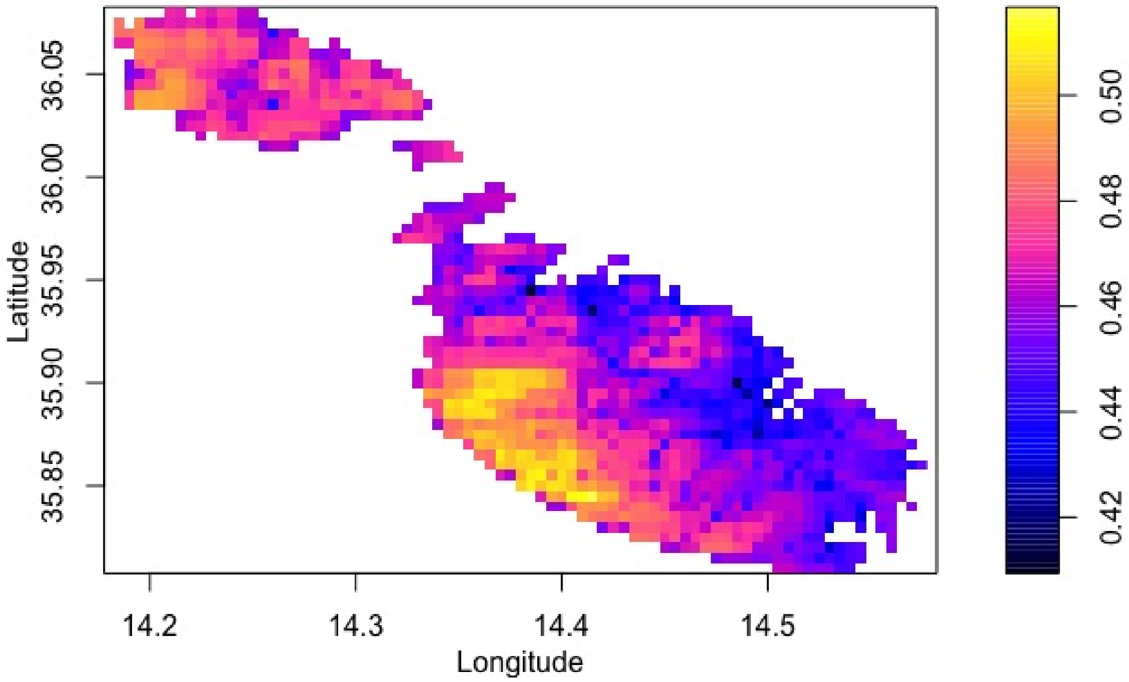


Fig. 4.17 Prediction uncertainties in the LTA regression model.

Chapter 5

Conclusion

5.1 Dissertation summary and key findings

In this dissertation, it was seen how spatio-temporal models are efficient tools to analyze data from air pollution monitoring. The introductory chapter included a brief overview of air pollution and the current situation in Malta. An outline of spatial statistics was also given together with a history of the main developments of spatial statistical models in air pollution monitoring. The monitoring of air pollution in Malta through the use of fixed stations and passive diffusion tubes was described in Chapter 2. This chapter also introduced the dataset, which consisted of NO_2 concentrations measured at 99 monitoring locations during a 3-year period. Geographic covariates such as elevation, population density, distances to roads, and industrial areas were added to the dataset to explain the spatial and temporal variations of the NO_2 concentrations.

Two spatial models were considered in this dissertation, and these were theoretically outlined in Chapter 3. The first model used a Universal Kriging structure to interpolate concentrations at unobserved locations and/or times. The second model consisted of a space-time mean field that incorporated dependence on geographic covariates together with seasonal and long-term trends; and a residual field having a spatial correlation structure. This model used smoothed orthogonal basis functions to capture the temporal variability in the data. In all models, parameter estimation was performed using maximum likelihood, and conditional expectations and reconstruction plots were reported. The predictive accuracy of both models was evaluated using cross-validation.

The two models were applied to the NO_2 dataset in Chapter 4. The spatio-temporal model was analyzed first. Two temporal basis functions were considered in the model. An independent and identically distributed covariance function was used to model the latent β-fields, whilst an exponential covariance function was used to model the ε-field. Suitable LUR covariates were chosen for the different β-fields. Assessing the model's predictive ability, the cross-validated R^2 value of approximately 0.55 suggests room for improvement in the spatio-temporal model..

The UK model was used to analyze the long-term average (LTA) of NO_2 at each monitoring location. First, a linear LTA regression model was constructed using elevation, distance to trunk, primary and secondary roads as explanatory variables. By comparing the raw observations together with the predicted observations, the LTA regression model provided a reasonably good fit. However, the variogram of the residuals showed no spatial dependence. For comparison purposes, a UK LTA model was constructed. Assessing the predictive ability of the UK model, the cross-validated R^2 value of approximately 0.52 is an indication for further improvement in the UK model.

Reconstructions of NO_2 across Malta identified interesting seasonal and spatial patterns of the pollutant. Evaluating the reconstructions from the spatio-temporal model, higher NO_2 concentrations were noted during winter. Furthermore, Gozo generated lower levels of NO_2 concentrations than in Malta. From the LTA reconstructions, higher predictions were seen in areas characterized mostly by trunk and primary roads, whilst lower predictions were seen in areas characterized mostly by secondary roads.

5.2 Limitations of the research

Due to the restriction of time and resources, the research conducted in this dissertation had some limitations. First and foremost, the geographic covariates did not help to establish a spatial effect in both the UK LTA model and the spatio-temporal model. Some of the covariates were transformed (such as taking truncated distances), however this still did not help in identifying spatial variability. If there was more time, other transformations would have been considered. Furthermore, problems were encountered when encoding the model. Some simulations with regards to parameter estimation and cross-validation took considerable time to generate a result.

5.3 Recommendations for future studies

In this dissertation, only the NO_2 pollutant was analyzed. Other air pollutants could be considered in the future such as particulate matter (PM) and volatile organic compounds (VOC's). These pollutants are present in motor vehicle emissions and industry related activities, which are major sources of air pollution in Malta.

Another recommendation would be to analyze air pollution using mobile stations. There are limitations in Malta since only one mobile station exists (Balzan 2012). However, adding this station to the other fixed stations would serve as a useful study to highlight any differences between the two types of stations.

Other adjustments of the spatio-temporal model could be made such as choosing different model specifications for the β-field and the ε-field. Furthermore, other transformations of the LUR covariates could be added to the model. All of these adjustments might help to capture any spatial and temporal variability in the data. Finally, a further improvement could be induced by grouping neighbouring stations and diffusion tubes together and performing the same analysis.

Bibliography

Amini, H., Taghavi Shahri, S. M., Henderson, S., Hosseini, V., Hassankhany, H., Naderi, M., Ahadi, S., Schindler, C., Künzli, N. & Yunesian, M. (2016). Annual and seasonal spatial models for nitrogen oxides in Tehran, Iran. Scientific Reports. 6.

Balzan, J. (2012). Mepa acquires new air monitoring mobile station. Malta Today. Retrieved February 3, 2018, from https://www.maltatoday.com.mt/news/national/21891/mepa-acquires-new-air-monitoring-mobile-station-20121016#.WxlcvzMza9Y

Becker, J. J.; Sandwell, D. T.; Smith, W. H. F.; Braud, J.; Binder, B.; Depner, J.; Fabre, D.; Factor, J.; Ingalls, S.; Kim, S. H.; Ladner, R.; Marks, K.; Nelson, S.; Pharaoh, A.; Sharman, G.; Trimmer, R.; VonRosenburg, J.; Wallace, G. & Weatherall, P. (2009). Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30PLUS Marine Geology Taylor & Francis, 32, 355-371

Camilleri, R. (2013). Nitrogen dioxide in the atmosphere : a study on the distribution of the air pollutant in the Maltese Islands (Master's thesis). University of Malta, Malta.

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., & Wang, N. (1997). Ozone Exposure and Population Density in Harris County, Texas. Journal of the American Statistical Association, 92(438), 392-404.

Chen, R., Samoli, E., Wong, C. M., Huang, W., Wang, Z., Chen, B. & Kan, H. (2012). Associations between short-term exposure to nitrogen dioxide and mortality in 17 Chinese cities: The China Air Pollution and Health Effects Study (CAPES), Environment International, Volume 45: 32-38,

Chiusolo, M., Cadum, E., Stafoggia, M., Galassi, C., Berti, G. et al. (2011) Short-term effects of nitrogen dioxide on mortality and susceptibility factors in ten Italian cities: the EpiAir Study. Environ Health Persp 119: 1233–1238

Cocchi, D., Greco, F., & Trivisano, C. (2007). Hierarchical space-time modelling of PM10 pollution. Atmos.Environ. 41, 532–542.

Cressie, N. A. C. (1993). Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

Cressie, N. et al. (1999) Spatial Analysis of Particulate Matter in an Urban Environment. In: Gómez-Hernández J., Soares A., Froidevaux R. (eds) geoENV II — Geostatistics for Environmental Applications. Quantitative Geology and Geostatistics, vol 10. Springer, Dordrecht

EEA (European Environmental Agency) (2017) Air quality in Europe — 2017. Copenhagen: EEA.

ERA (Environment and Resource Authority) Data from Air Monitoring Stations. Retrieved 20 January 2018 from https://era.org.mt/en/Pages/Data-from-Air-Monitoring-Stations.aspx

Guttorp, P., Meiring, W. & Sampson, P. (1994). A space-time analysis of ground-level ozone data. Environmetrics. 5. 241-254.

Haas, T. C. (1995), Local prediction of a spatio-temporal process with an application to wet sulfate deposition. J. Amer. Statist. Assoc. 90, 1189-1199

Kibria, B. M. G., Sun, L., Zidek, J., & Le, N. (2002). Bayesian Spatial Prediction of Random Space-Time Fields with Application to Mapping PM2.5 Exposure. Journal of the American Statistical Association, 97(457), 112-124.

Lindström, J. (2017). Spatial Statistics with Image Analysis: Lecture 4 [Powerpoint slides]. Retrieved 21 January 2018 from http://www.maths.lth.se/matstat/kurser/fmsn20masm25/material_ht17/F03-1x3.pdf

Lindström, J., Szpiro, A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2013a). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environmental and Ecological Statistics, 21(3), 411-433.

Lindstrom J, Szpiro A., Sampson P. D., Bergen S., & Oron A. P. (2013b). SpatioTemporal: SpatioTemporal Model Estimation. R package version 1.1.7, Retrieved from http://CRAN.R-project.org/package=SpatioTemporal

Madany, I. M., & Danish, S. (1993). Spatial and temporal patterns in nitrogen dioxide concentrations in a hot desert region. Atmospheric Environment. Part A. General Topics, 27(15), 2385-2391.

MITA (Ministry for infrastructure, Transport and Communication) (2009) Air Quality Plan. Retrieved from http://www.transport.gov.mt/admin/uploads/media-library/files/MITC%201_Air%20Quality%20Plan%20v2.pdf

Nash, D.G. & Leith, D. (2010). Use of Passive Diffusion Tubes to Monitor Air Pollutants, Journal of the Air & Waste Management Association, 60:2, 204-209,

NSO (National Statistics Office) (2011a) Census of Population and Housing 2011. Valletta: NSO.

NSO (National Statistics Office) (2011b) The Climate of Malta: statistics, trends and analysis, 1951-2010.. Valletta: NSO.

NSO (National Statistics Office) (2016) Annual Report 2016. Valletta: NSO.

NSO (National Statistics Office) (2018) Sea transport between Malta and Gozo: Q4/2017. Valletta: NSO.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Ribeiro Jr., P.J. & Diggle, P.J. (2001). geoR: A package for geostatistical analysis. R-NEWS, Vol 1, No 2, 15-18.

Sahu, S. K. and Bakar, K. S. (2011). A comparison of Bayesian Models for Daily Ozone Concentration Levels Statistical Methodology, DOI: 10.1016/j.stamet.2011.04.009.

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D. & Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics, 21: 606-631.

United Nations. (1997). Glossary of Environmental Statistics, Studies in Methods (Series F No. 67). Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=1558

Wessel, P., and Smith, W. H. F. (1996). A Global Self-consistent, Hierarchical, High-resolution Shoreline Database, J. Geophys. Res., 101, 8741-8743.

WHO (World Health Organization) (2016). Ambient air pollution: A global assessment of exposure and burden of disease. Retrieved from http://www.who.int/phe/health_topics/outdoorair/global-exposure-assessment-faq/en/

Zammit, L.C. (2013). Spatio-Temporal Models for Traffic and Air Pollution. Study presented at IET Present around the World Competition, St Julan's Malta

Zammit, L.C., Scerri, K., Attard, M., Bajada, T. & Scerri, M. (2011). Spatio-Temporal Analysis of Air Pollution Data in Malta. Paper presented at 11th International Conference of GeoComputation 2011, organised by the GeoComputation Community, University College London, UK, 20th – 22nd July.

Zhang, Y. D., Wang, S. & Ji, G. (2013). A Rule-Based Model for Bankruptcy Prediction Based on an Improved Genetic Ant Colony Algorithm. Mathematical Problems in Engineering. 2013. 1-10.

Zimmerman, D. L. & Stein, M. (2010). Classical Geostatistical Methods. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, Handbook of Spatial Statistics, pages 29–44. Chapman & Hall/CRC.

Appendix

A1 The districts of Malta

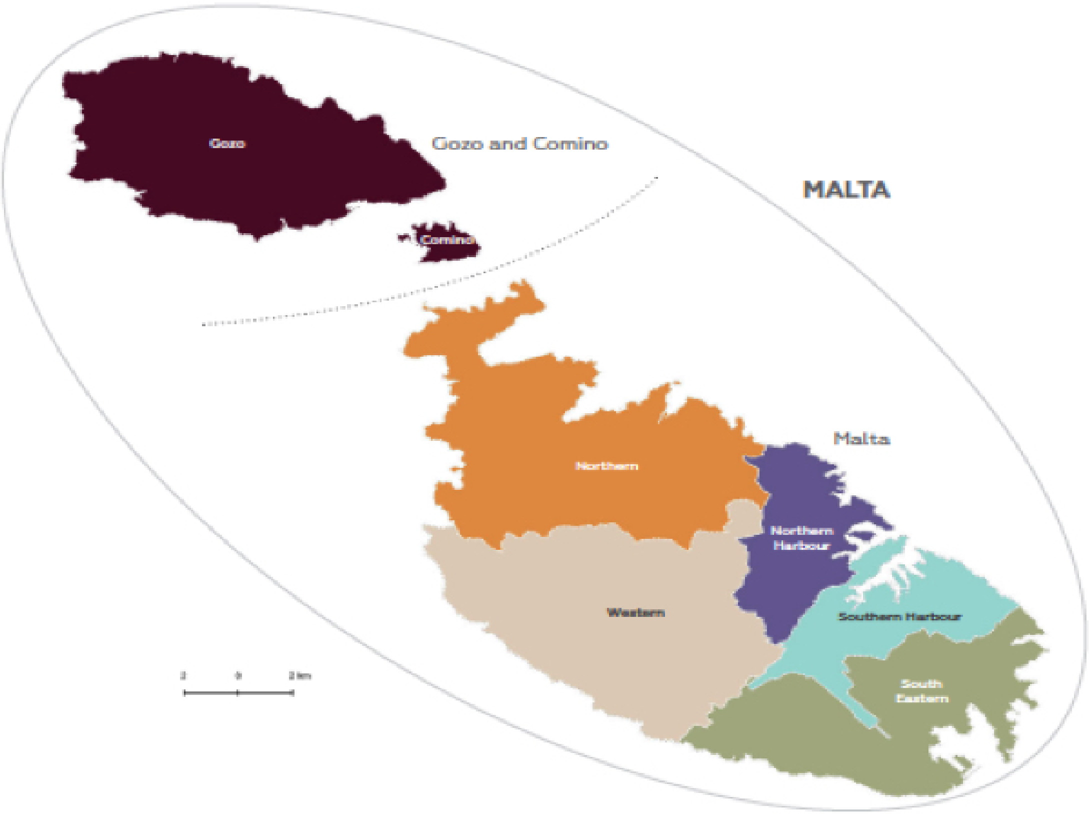


Fig. A1 The 6 districts of Malta are divided according to the Nomenclature of Territorial Units for Statistics (NUTS) classification. This is a hierarchical classification which divides the economic territory of the European Union for the purpose of producing regional statistics that are comparable across the European Union. Source (NSO 2016)

A2 The localities of Malta

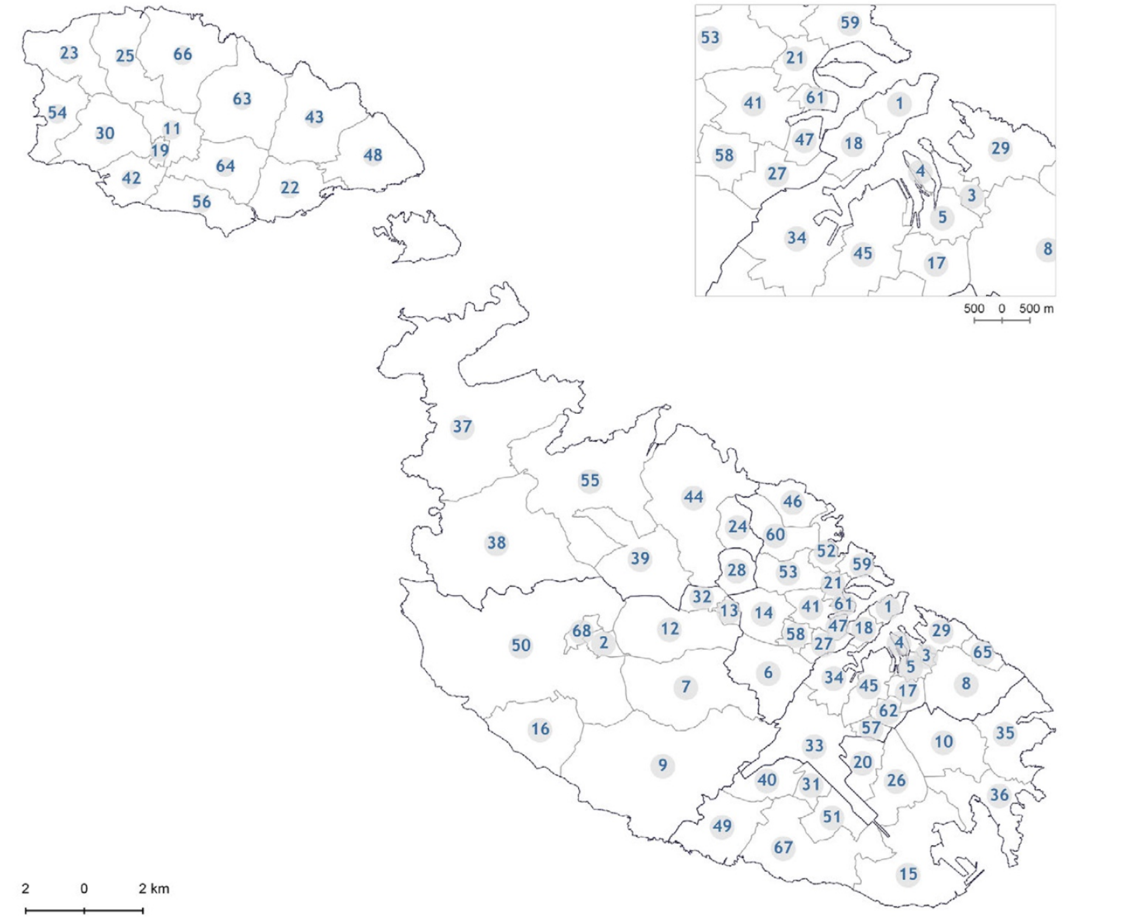


Fig. A2 A map of Malta divided into 68 localities of Malta in graphical form. The names of the numbered localities on the map are given in Table A1. Source: NSO (2016)

Number	Locality	Number	Locality
1	Valletta	35	Marsaskala
2	Mdina	36	Marsaxlokk
3	Vittoriosa	37	Mellieħa
4	Senglea	38	Mġarr
5	Cospicua	39	Mosta
6	Ħal Qormi	40	Mqabba
7	Ħaż-Żebbug	41	Msida
8	Ħaż-Żabbar	42	Munxar
9	Siġġiewi	43	Nadur
10	Żejtun	44	Naxxar
11	Victoria	45	Paola
12	Ħ’Attard	46	Pembroke
13	Balzan	47	Tal-Pieta’
14	Birkirkara	48	Qala
15	Birżebbuġa	49	Qrendi
16	Ħad Dingli	50	Rabat (Malta)
17	Fgura	51	Ħal Safi
18	Floriana	52	St Julian’s
19	Fontana	53	San Ġwann
20	Gudja	54	San Lawrenz
21	Gżira	55	St Paul’s Bay
22	Għajnsielem and Comino	56	Ta’ Sannat
23	Għarb	57	Santa Luċija
24	Ħal Għargħur	58	Santa Venera
25	Għasri	59	Tas-Sliema
26	Ħal Għaxaq	60	Swieqi
27	Ħamrun	61	Ta’ Xbiex
28	Iklin	62	Ħal Tarxien
29	Kalkara	63	Xagħra
30	Ta’ Kerċem	64	Xewkija
31	Ħal Kirkop	65	Xgħajra
32	Ħal Lija	66	Żebbug (Gozo)
33	Ħal Luqa	67	Żurrieq
34	Marsa	68	Mtarfa

Table A1 The names of the 68 localities in Malta corresponding to Figure A2. Source: NSO (2016)

A3 Prediction of β-fields and unobserved time points

The following predictions are presented in Lindström et al. (2013b). Given the estimated parameters, the predictions of β-field are given by the conditional expectation of B_u given Ψ and Y:

$$\mathbb{E}(B_u|Y,\Psi) = X_u\hat{\alpha} + \Sigma_{B,uo}F^T\tilde{\Sigma}^{-1}(Y - \tilde{X}\hat{\alpha}) \tag{.1}$$

The corresponding conditional variance of β_u given Ψ, α and Y is:

$$\mathbb{V}(B_u|Y,\Psi,\alpha) = \Sigma_{B,uu} - \Sigma_{B,uo}F^T\tilde{\Sigma}^{-1}F\Sigma_{B|Y}\Sigma_{B,ou} \tag{.2}$$

Adding the uncertainty in the regression coefficients, (.2) becomes:

$$\begin{aligned}\mathbb{V}(B_u|Y,\Psi) =& \mathbb{V}(B_u|Y,\psi,\alpha) + ([0\ \ X_u] + \Sigma_{B_{uo}}F^T\tilde{\Sigma}_v^{-1}\tilde{X})(\tilde{X}^T\tilde{\Sigma}^{-1}\tilde{X})^{-1}\\ &([0\ \ X_u] + \Sigma_{B_{uo}}F^T\tilde{\Sigma}^{-1}\tilde{X})^T\end{aligned} \tag{.3}$$

Let $y(s,t_u)$ denote an unobserved time point. The conditional expectation of $y(s,t_u)$ given Y and Ψ is:

$$\mathbb{E}(y(s,t_u|Y,\Psi) = F_u\mathbb{E}(B_u|,Y,\Psi) \tag{.4}$$

The corresponding conditional variance of $y(s,t_u)$ given Y, Ψ and α is:

$$\mathbb{V}(y(s,t_u)|Y,\Psi) = F_u\mathbb{V}(B_u|Y,\Psi,\alpha)F_u^T + \Sigma_{v,uu} \tag{.5}$$

A4 Parameter estimates for the spatio-temporal model

Field	Parameter	Estimated value
β_0	Intercept	5.63
	Elevation	0.01
	Latitude	-1.41
	Distance to trunk road	-0.19
	Distance to primary road	-0.15
	Distance to secondary road	-0.11
	τ_0^2	-1.69
β_1	Intercept	0.28
	Elevation	0.01
	Population density	-0.02
	Distance to trunk road	-0.01
	Distance to primary road	-0.01
	Distance to secondary road	-0.01
	Distance to industrial road	0.01
	Distance to coast road	-0.01
	τ_1^2	-6.47
β_2	Intercept	-26.40
	Latitude	0.50
	Longitude	0.57
	Elevation	0.02
	Population density	0.03
	Distance to coast	0.01
	τ_2^2	-6.79
ε	ϕ_ε	-2.22
	σ_ε^2	-5.05
	τ_ε^2	-4.85

Table A2 Parameter estimates for the spatio-temporal model.

Master’s Theses in Mathematical Sciences 2018:E47
ISSN 1404-6342

LUNFMS-3077-2018

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/