Spatio-temporal modelling for non-stationary point referenced data

by

Lindsay Robert Morris

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Statistics.

Victoria University of Wellington
2021

Abstract

Spatial and spatio-temporal phenomena are commonly modelled as Gaussian processes via the geostatistical model (Gelfand & Banerjee, 2017). In the geostatistical model the spatial dependence structure is modelled using covariance functions. Most commonly, the covariance functions impose an assumption of spatial stationarity on the process. That means the covariance between observations at particular locations depends only on the distance between the locations (Banerjee et al., 2014). It has been widely recognized that most, if not all, processes manifest spatially non-stationary covariance structure Sampson (2014). If the study domain is small in area or there is not enough data to justify more complicated non-stationary approaches, then stationarity may be assumed for the sake of mathematical convenience (Fouedjio, 2017). However, relationships between variables can vary significantly over space, and a 'global' estimate of the relationships may obscure interesting geographical phenomena (Brunsdon et al., 1996; Fouedjio, 2017; Sampson & Guttorp, 1992).

In this thesis, we considered three non-parametric approaches to flexibly account for non-stationarity in both spatial and spatio-temporal processes. First, we proposed partitioning the spatial domain into sub-regions using the K-means clustering algorithm based on a set of appropriate geographic features. This allowed for fitting separate stationary covariance functions to the smaller sub-regions to account for local differences in covariance across the study region. Secondly, we extended the concept of covariance network regression to model the covariance matrix of both spatial and spatio-temporal processes. The resulting covariance estimates were found to be more flexible in accounting for spatial autocorrelation than

standard stationary approaches. The third approach involved geographic random forest methodology using a neighbourhood structure for each location constructed through clustering. We found that clustering based on geographic measures such as longitude and latitude ensured that observations that were too far away to have any influence on the observations near the locations where a local random forest was fitted were not selected to form the neighbourhood.

In addition to developing flexible methods to account for non-stationarity, we developed a pivotal discrepancy measure approach for goodness-of-fit testing of spatio-temporal geostatistical models. We found that partitioning the pivotal discrepancy measures increased the power of the test.

Acknowledgements

The last three years and three months have simultaneously felt very short and incredibly long. During that time, I realised that the task of writing a thesis, from planning to producing, was no easy feat. I cannot express enough just how much of a bumpy ride undertaking a PhD can be. I am certain I would not have made it to submission without my incredible support system.

I would like to express my deepest appreciation to my supervisor, Dr Nokuthaba Sibanda. Nokuthaba has been my source of academic support for many years. Her guidance and overflowing knowledge has helped me countless times. I have always been able to depend on her kindness and experience throughout my PhD. I hope one day to be half the statistician Nokuthaba is.

I acknowledge the support from the administrative staff at the School of Mathematics and Statistics, particularly Alec, Caitlin, Ginny, and Simonette. There have been many times when I have felt overwhelmed by administration and these four have always managed to alleviate the stress.

I am also thankful for my office mates, Akib, Kien, and Lingyu, who I have had the pleasure to share the journey to good research with.

I would like to extend my sincere thanks to my dad, Craig, and my younger

siblings, Ryan, Brooke, and Bree. I am grateful for their support and understanding, especially during the stressful months.

I also gratefully acknowledge Karen & Ash, and Craig for their continued support.

To my friends Sam & Steven, Tash & Tapiwa, Katie, Laura, and Karina, I extend my deepest gratitude. I am so thankful for the many coffee dates, lunch breaks, road trips, and dinner parties. The love and support I have received throughout has been unwavering.

I express my utmost gratitude to Charlotte Page. Charlotte and I have been through every stage of university together, from studying chemistry in undergrad, to writing theses for PhD. She has been a source of light and laughter at Victoria University and I am grateful that we got to share the highest of highs and the lowest of lows together.

To the love of my life, Bradley Pratt. I cannot begin to describe how grateful I am for your love and support over the last three years. You have played a central role throughout my PhD. Thank you for listening, for your patience, and for providing me with advice, comfort, and care. I certainly believe that being the partner of a PhD student is harder than being a PhD student.

Dedication

To my mum, Christina. Your love and support has been felt at every step throughout my life, and the last three years and three months were no exception. Thank you for your encouragement, humour, and for keeping me grounded. Thank you for the daily texts, phone calls, and messages of comfort. Thank you for always reminding me that Nana would be proud. You're the best mum ever. I promise I'll go and get a job now.

I dedicate my thesis to you, Mum.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last few decades have seen increasingly rapid advances in the field of spatial and spatio-temporal statistics. The procedure for describing spatial variation has evolved from what was once considered *ad hoc* to that, which is based on models (Gelfand & Banerjee, 2017). The application of spatial and spatio-temporal statistical methods is diverse. It has use in climatology, ecological and environmental sciences, the health sector, real estate marketing, demography, and mining.

Spatial data are often described as one of three types: areal, point pattern, or point referenced. Areal (or lattice) data describe measurements that have been observed for a finite number of areal units with well-defined boundaries (Banerjee et al., 2014). The data are summaries defined on a regular or irregular lattice. An example of areal data over a regular lattice are agricultural field trials where the plots cultivated are arranged regularly. More commonly, however, areal data are arranged over irregular lattices such as regional boundaries in a country or other geographical areas. A visualisation of areal data is given in Figure 1.1. Here, the estimated percentage of the population of New Zealand that have type 1 and 2 diabetes is projected on a 2-dimensional map of the country. The percentages are averaged over each District Health Board (DHB) region with a single value representing the entire areal unit. From the diagram, we can infer

that the percentage of type 1 and 2 diabetes is higher for DHB regions in the north compared to the south. Models for areal data are typically seen

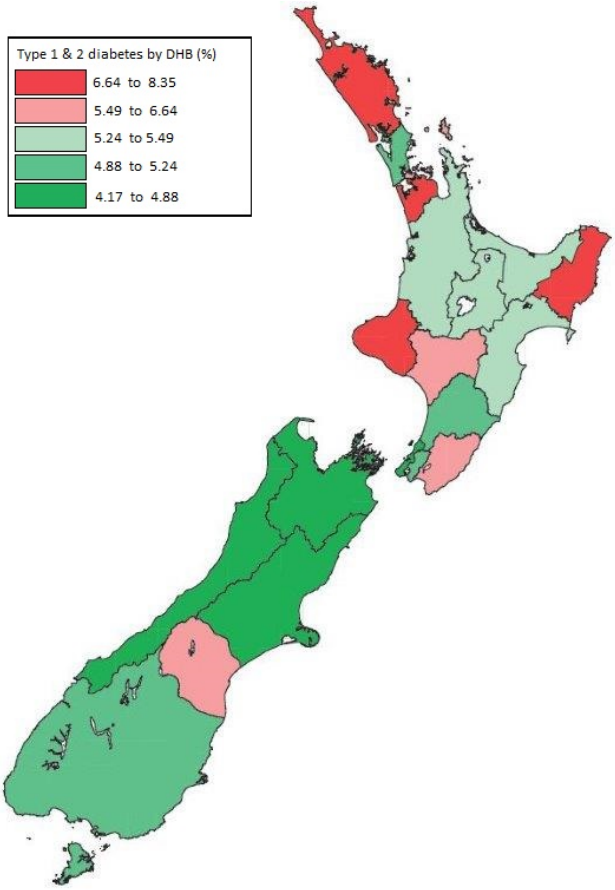


Figure 1.1: Estimated percentage of New Zealand population that has type 1 and 2 diabetes by district health board area using data from the New Zealand Society for the Study of Diabetes as at 31 December 2013 (Parliamentary Library, NZ, 2014).

as being too restrictive in the range of spatial dependence they can successfully model, but offer the advantage of being more computationally efficient (White & Ghosh, 2009). The most common model fit to areal data is the conditional autoregressive (CAR) model, which makes the assumption that neighbouring areal units are more likely to be correlated than

non-neighbouring areal units.

Another type of spatial data are point pattern data. Point pattern data describe the situation where the locations of random events are considered random (Banerjee et al., 2014). Of interest is the location and pattern of observations, and not the value of the observation itself. Figure 1.2 provides an illustration of how to visualise point pattern data. The 'quakes' dataset in R gives the locations of shallow earthquakes across New Zealand and the Pacific. These are plotted on a map and allow informal inferences to be made about the pattern of occurrence. From the plot, a clear pattern emerges that correlates to the Ring of Fire where a series of volcanic eruptions and earthquakes frequently occur. Although the magnitude of the earthquake is reported as well, from a point pattern perspective it is not needed. Point pattern data are often modelled using Poisson processes.

The third type of spatial data are point referenced data. In this thesis, we focused on spatial and spatio-temporal methods for point referenced data. This type of data (often referred to as geostatistical data) describe measurements that have been observed at a particular fixed location. Formally, $\boldsymbol{Y}(\boldsymbol{s})$ is a random vector at a location $\boldsymbol{s} \in \mathbb{R}^r$, where $\boldsymbol{s}$ varies continuously over D, a fixed subset of $\mathbb{R}^r$ that contains an r-dimensional rectangle of positive volume (Banerjee et al., 2014). Point referenced data can be visualised on a 2-dimensional or 3-dimensional map. Figure 1.3 displays median nitrate levels in groundwater across New Zealand from 1995 to 2006. This plot allows us to notice clusters of observations with relatively similar values, which is a sign of spatial autocorrelation. We describe this concept in the next section. Another example of point referenced data is given by Figure 1.4. It shows the catch weight in kilograms per square kilometre of the fish species hoki caught in 1255 trawls of the sub-Antarctic region. We describe the hoki data in more detail in Section 2.7.2. The advantage of working with point referenced data is that it provides the most information of the three spatial data types. Not only does this type of data contain an observable value, it also gives the exact location. Models for

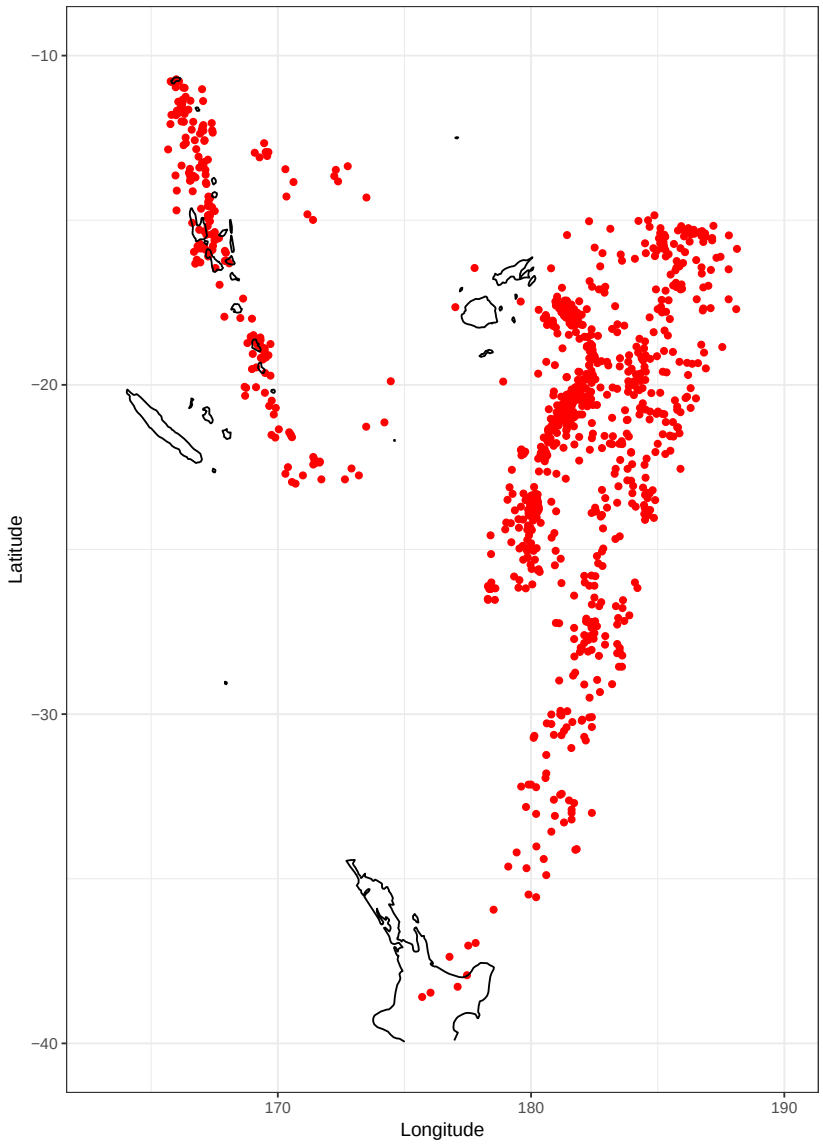


Figure 1.2: The Pacific experiences large numbers of small earthquakes, in well-defined belts stretching across the Pacific Islands to New Zealand. This pattern is part of the 'Ring of Fire', the almost continuous belt of volcanoes and earthquakes rimming the Pacific Ocean.

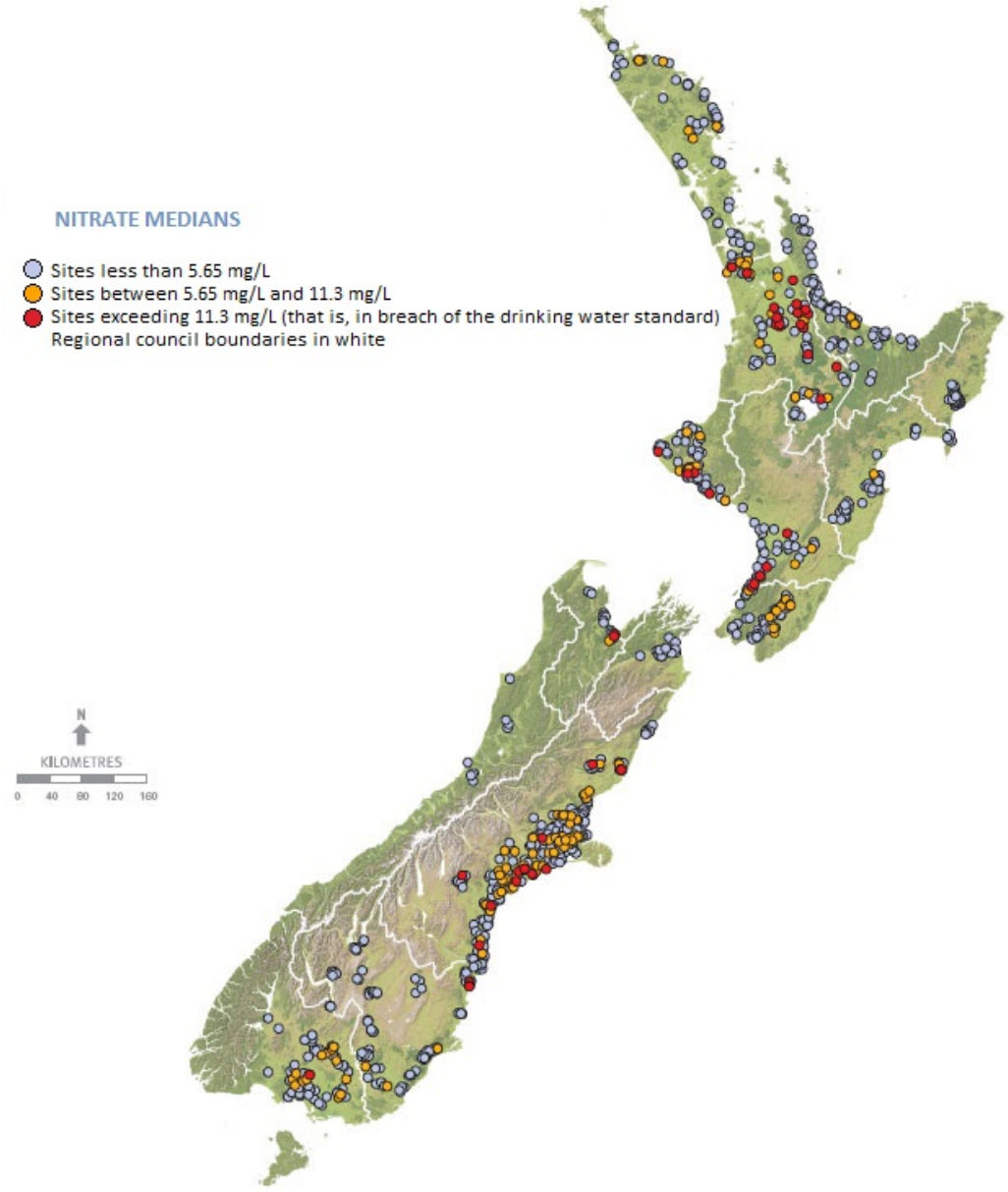


Figure 1.3: Median nitrate levels in groundwater, 1995 – 2006. Map showing all New Zealand monitoring wells colour-coded according to the nitrate category they fall into (3 categories covering the range 0 to >11.3 mg/L). Regions with a significant proportion of wells that have median nitrate exceeding 5.65 mg/L (i.e. half the New Zealand Drinking Water Standard) are Waikato, Manawatu, Wairarapa, Taranaki, Canterbury and Southland (Ministry for the Environment & Statistics New Zealand, 2007).

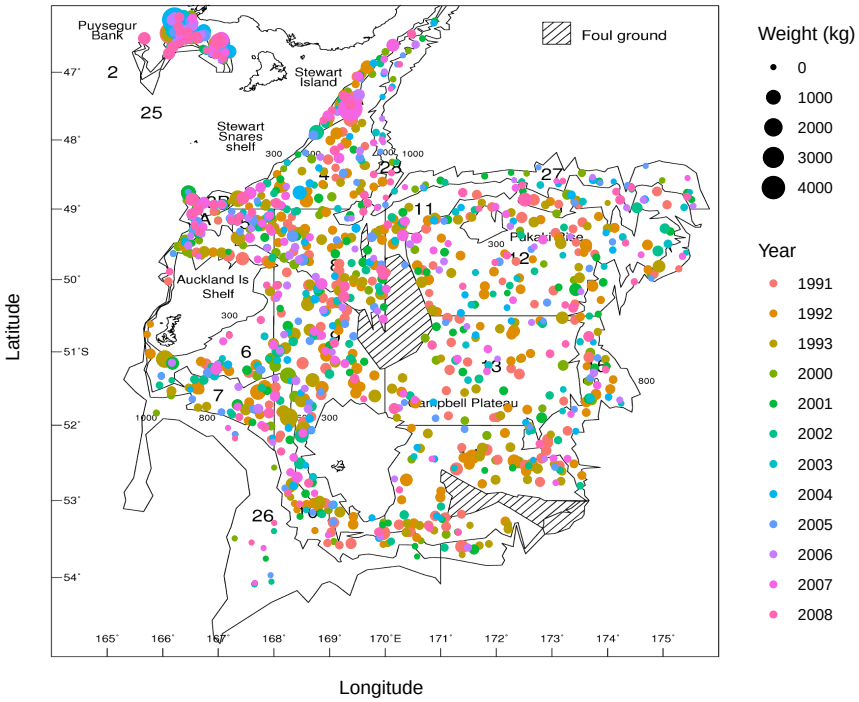


Figure 1.4: Map of 1255 trawls where hoki was caught in the sub-Antarctic region. The size of a point is representative of the catch weight in kg, whereas, the colour is representative of the year (Morris, 2017).

point referenced data can be computationally intensive (Gelfand & Banerjee, 2017; White & Ghosh, 2009; Cameletti et al., 2013), however, they offer more flexibility in capturing the spatial dependence when a large number of observations are measured at different sites (White & Ghosh, 2009).
There are several main objectives for modelling point referenced data. Researchers might be interested in: identifying and estimating the effects of predictors of a point referenced variable, describing and accounting for the covariance structure of a point referenced variable, or predicting and interpolating point referenced variables at locations not sampled. Because of the nature of spatial data, care needs to be taken to account for spatial dependence, spatial autocorrelation, and other spatial phenomena when modelling point referenced data. We introduce these concepts in the following sections.

1.1 Autocorrelation

An important concept to consider when fitting models to spatial data is that of Autocorrelation. Autocorrelation was defined by Yule (1921) as the dependence of successive observations of a single variable. An assumption of independence between observations is often made when constructing a model. When autocorrelation exists within data, care must be taken to account for such dependencies. There are different types of autocorrelation, with the most common being temporal and spatial autocorrelation.
Temporal autocorrelation is the correlation between a set of observations of a single variable observed at different time points. A simple way to account for this type of autocorrelation is to assume that observations that were observed closer in time are more correlated than those separated by a large time distance. There are many models that can account for temporal autocorrelation, with one being the autoregressive (AR) model.
Spatial autocorrelation is the correlation between a set of observations of a single variable observed at different locations or in different areas. The

most prominent law in geography is Tobler's first law, and states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). It means that characteristics of phenomena at one location on Earth tend to be similar to those at nearby locations. Throughout this thesis, we will see how this concept can be incorporated into models to account for spatial autocorrelation.

A model that does not appropriately consider spatial autocorrelation is misspecified and failure to account for it can result in biased predictions and inflated errors. Proper specification requires that any spatial association is accounted for within the model properly (Elhorst et al., 2010). In this thesis, we we propose several modelling approaches that allow and account for spatial autocorrelation within point referenced data.

1.2 Non-stationarity

Another important concept that needs to be considered when modelling spatial data is non-stationarity. We describe this concept in greater detail in Chapter 2. Fundamentally, models for spatial data take into account Tobler's first law of geography (Tobler, 1970), which states that "everything is related to everything else, but near things are more related than distant things." For point referenced data, this is often carried out by fitting a model defined by a particular covariance structure. The covariance structure is commonly estimated assuming that the covariance between any pair of point referenced observations, separated by a distance d, is the same, regardless of the location, and this is known as stationarity. However, it is more likely for the covariance structure to vary with spatial location, and this is known as non-stationarity.

In this thesis, we proposed several original methodologies to take into account the concepts of spatial autocorrelation and non-stationarity. The following sections detail the aim and outline of this thesis.

1.3 Aim of thesis

The main aim of this thesis was to develop a range of new methodologies that account for non-stationarity in spatial and spatio-temporal point referenced data. We developed three distinct methodologies, each with a non-parametric component. We first contributed to the literature on geostatistical models, with a partitioning approach based on the K-means clustering algorithm. The intention was to partition the spatial domain of point referenced data and assume stationarity within the sub-regions. We then contributed to the geostatistical modelling literature an approach based on covariance regression network models. By applying a technique used in the network analysis literature, we proposed a more flexible covariance function based on regression of the covariance matrix using an network structure estimated from the locations of the point referenced data. We then took a non-parametric, machine learning approach, and proposed a modification of geographic random forests that involved constructing a neighbourhood structure based on K-means clustering. We made a further contribution by developing geographic random forest approach for spatio-temporal point referenced data.
A secondary aim of this thesis was to develop a goodness-of-fit test for Bayesian spatial and spatio-temporal geostatistical models. The literature on goodness-of-fit tests for geostatistical models is surprisingly sparse, and we made a contribution to it in the form of pivotal discrepancy measures. We now outline the layout of this thesis.

1.4 Outline of thesis

This thesis is organized into seven chapters, with the main contributions detailed in Chapters 3, 4, 5, and 6.
In Chapter 2, we introduced several key concepts and methodologies that are central to this thesis. The chapter includes topics on stochastic pro-

cesses, stationarity, detecting spatial autocorrelation and non-stationarity, the geostatistical model, Bayesian methods, and model assessment tools. In addition, we introduced two datasets that were used as case studies in the main chapters. The first set of data detail information from monitoring stations placed at 40 fixed locations throughout New Zealand that measured the concentration of particulate matter in 2013. The second set of data contains information from research trawl surveys of the sub-Antarctic region that were carried out by the National Institute of Water and Atmospheric Research (NIWA) for the Ministry for Primary Industries, New Zealand (MPI), from 1991 to 1993, and 2000 to 2008.

In Chapter 3, we presented a geostatistical model for spatial and spatio-temporal point referenced data that used partitioning via the K-means algorithm to account for non-stationarity. This was followed by Chapter 4, in which we presented a covarianace regression network modelling approach for spatial and spatio-temporal point referenced data. We then presented a non-parametric geographic random forest approach for making predictions from spatial and spatio-temporal point referenced data in Chapter 5. Following in Chapter 6, we presented a pivotal discrepancy measure for goodness-of-fit testing of geostatistical models fitted to spatio-temporal data in a Bayesian context.

In Chapters 3 to 6, we presented a series of simulation studies that were conducted to evaluate the performance of our proposed methodologies. Further, the methods were applied to the particulate matter data and hoki data to test their viability in reality.

We concluded the thesis in Chapter 7, with a discussion of the proposed methodologies, their limitations, and future research considerations.

Chapter 2

Preliminary methodology

In this chapter, we introduced key methodology used and referenced to in this thesis. We begin with a section dedicated to stochastic processes within the spatial statistics context. This is followed by sections that outline procedures for detecting spatial autocorrelation and non-stationarity. We then introduced the geostatistical model and the Bayesian framework. A short introduction is given to model comparison and assessment tools. We conclude the chapter with descriptions of two sets of data that we used in case studies throughout the thesis.

2.1 Spatial and spatio-temporal stochastic processes

Within the point referenced data setting, modelling is carried out by specifying random surfaces over $\mathbb{R}^2$. One way to do this is to model the surface as a realization of a stochastic process (Gelfand & Schliep, 2016). In this section, we introduced stochastic processes, particularly spatial and spatio-temporal stochastic processes, for modelling point referenced data.

2.1.1 Definition

A collection of point referenced observations from a potentially infinite number of measurements is a realization of a stochastic process. A stochastic process is a collection of random variables, $Y(\boldsymbol{s};\omega) \equiv \{Y(\boldsymbol{s};\omega) : \boldsymbol{s} \in D;\omega \in \Omega\}$, on some probability space (Ω, F, P) indexed by a variable $\boldsymbol{s} \in D$ where D is a fixed subset of $\mathbb{R}^d$ with positive d-dimensional volume (Cressie, 1993; Billingsley, 2013). The realization of the process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ would correspond to a particular value of ω, say ω_0. For this thesis, we suppress the dependence of $Y(.)$ on $\omega \in \Omega$, and define the stochastic process as,

$$Y(\boldsymbol{s}) \equiv \{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}. \tag{2.1}$$

The stochastic process is called a time series when $\boldsymbol{s}$ represents continuous or discrete indices in time, t. When $\boldsymbol{s}$ represents spatial locations that are (usually) fixed over a continuous d-dimensional set ($D \subseteq \mathbb{R}^d$), where $d = 2$ or $d = 3$, then the process is considered spatial stochastic (Cressie, 1993). Furthermore, $Y(\boldsymbol{s},t) \equiv \{y(\boldsymbol{s},t) : (\boldsymbol{s},t) \in D\}$ is called a spatio-temporal stochastic process when $(\boldsymbol{s},t)$ represent spatial locations at time t fixed over a continuous d-dimensional set ($D \subseteq \mathbb{R}^{d-1} \times \mathbb{R}$) (Cameletti et al., 2013). Then $(Y(\boldsymbol{s_1},t),...,Y(\boldsymbol{s_n},t))'$ is a random vector, whose multivariate distribution reflects the spatial dependencies in the variable of interest.

The expectation of mean function, $\mu(.)$, of a stochastic process is defined by,

$$\mu(\boldsymbol{s}) = \mathrm{E}(Y(\boldsymbol{s})), \tag{2.2}$$

and the covariance function, $C(.,.)$, of a stochastic process is defined for any pair $(\boldsymbol{s}_i, \boldsymbol{s}_j)$,

$$\begin{aligned} C(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \mathrm{Cov}(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) \\ &= \mathrm{E}\big((Y(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i))(Y(\boldsymbol{s}_j) - \mu(\boldsymbol{s}_j))\big), \end{aligned} \tag{2.3}$$

(Paciorek, 2003).

Stochastic processes are often defined by their finite-dimensional distributions (Cressie, 1993). A Gaussian process is a stochastic process whose finite dimensional distributions are multivariate normal, and are completely specified by their mean and covariance functions, just as multivariate Gaussian distributions are specified by their mean vector and covariance matrix (Paciorek, 2003). The covariance function for a Gaussian process must be positive definite, in order to satisfy that the finite dimensional distributions are consistent (Stein, 2012). A covariance function is positive definite if it satisfies,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\boldsymbol{s_i}, \boldsymbol{s_j}) \geq 0, \tag{2.4}$$

for every n, every collection $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$, and any vector of real numbers $\boldsymbol{a}$. Gaussian processes are widely used in modelling spatial and spatio-temporal data (Paciorek, 2003). In most applications, an assumption of stationarity is made on the covariance function for simplicity. We define stationarity in the following section.

2.1.2 Stationarity

One of the main assumptions made when modelling a spatial or spatio-temporal process is that the process is stationary. A stochastic process is called strictly stationary if the finite dimensional joint distributions are invariant under translation of the spatial coordinates. In other words,

$$\Pr(Y(\boldsymbol{s_1} + \boldsymbol{h}) < y_1, ..., Y(\boldsymbol{s_n} + \boldsymbol{h}) < y_n) = \Pr(Y(\boldsymbol{s_1}) < y_1, ..., Y(\boldsymbol{s_n}) < y_n), \tag{2.5}$$

for all vectors $\boldsymbol{h} \in \mathbb{R}^d$ (Cressie, 1993; Gelfand et al., 2010; Schabenberger & Gotway, 2017). This type of stationarity is a strong condition. For most spatial statistical methods it is satisfactory to have stationarity conditions based on moments of the joint distributions rather than the distributions themselves.

Second-order stationarity is assumed so that the data are considered representative of a complete sampling of a single realization. Formally, and in general, the stochastic process, $Y(.)$, that satisfies

$$\mathrm{E}\big(Y(\boldsymbol{s})\big) = \mu, \forall \boldsymbol{s} \in D, \tag{2.6}$$

and

$$\mathrm{Cov}\big(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\big) = C(\boldsymbol{s}_i - \boldsymbol{s}_j), \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D, \tag{2.7}$$

is defined to be second-order stationary, where C is the covariance function. The mean of a second-order stationary spatial process is constant and the covariance between elements of $Y(\boldsymbol{s})$ is a function only on their spatial separation (represented by $\boldsymbol{h}$, named the lag vector) and illustrates the lack of importance of absolute coordinates.
The covariance function of a second-order stationary spatial stochastic process has several nice properties that are listed below (Schabenberger & Gotway, 2017):

1. $C(\mathbf{0}) \geq 0$;
2. $C(\boldsymbol{h}) = C(-\boldsymbol{h})$, i.e., C is an even function;
3. $C(\mathbf{0}) \geq |C(\boldsymbol{h})|$;
4. $C(\boldsymbol{h}) = \mathrm{Cov}\big(Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})\big) = \mathrm{Cov}\big(Y(\mathbf{0}), Y(\boldsymbol{h})\big)$;
5. If $C_j(\boldsymbol{h})$ are valid covariance functions, $j = 1, ..., k$, then $\sum_{j=1}^{k} b_j C_j(\boldsymbol{h})$ is a valid covariance function, if $b_j \geq 0 \; \forall j$;
6. If $C_j(\boldsymbol{h})$ are valid covariance functions, $j = 1, ..., k$, then $\prod_{j=1}^{k} C_j(\boldsymbol{h})$ is a valid covariance function;
7. If $C_j(\boldsymbol{h})$ are valid covariance functions, $j = 1, ..., k$, then $\sum_{j=1}^{k} b_j C_j(\boldsymbol{h})$ is a valid covariance function in $\mathbb{R}^d$, then it is also a valid covariance function in $\mathbb{R}^p$, $p < d$.

Properties 5 – 7 require that the covariance function of a stationary process be valid. A covariance function $C(\boldsymbol{h}) = C(\boldsymbol{s_i} - \boldsymbol{s_j})$ of a second-order stationary spatial stochastic process is considered valid if it is a symmetric and positive definite function that satisfies Equation 2.4 for every n, every collection $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$, and any vector of real numbers $\boldsymbol{a}$.

2.1.3 Non-stationarity

The assumption of stationarity is an exception, rather than a generality (Fouedjio, 2017). When modelling some phenomena, particularly environmental, the stationarity assumption may not be suitable (Blangiardo & Cameletti, 2015), (Sampson & Guttorp, 1992). That is, the covariance structure between observations across the domain might not be constant, since geographic variables might influence the spatial dependence structure. In these cases, it would be more suitable to consider non-stationary spatial and spatio-temporal methods.
The spatial and spatio-temporal modelling literature is rich with modelling approaches that involve estimating an explicit non-stationary covariance function. A non-stationary covariance function varies over the spatial and/or spatio-temporal domain. Paciorek & Schervish (2006) proposed a class of non-stationary covariance functions that can be constructed from a stationary covariance function. In this thesis, we considered non-parametric solutions to account for non-stationarity that do not require an explicit non-stationary covariance function. We present these in the chapters that follow.

2.2 Detecting spatial autocorrelation

2.2.1 Visualizing spatial autocorrelation

We can visualize spatial autocorrelation in a point referenced dependent variable by plotting it against the geographic coordinates at which the ob-

servations were made. To obtain a smoother surface, the observations can be interpolated using. for example, Akima interpolation (Akima, 1978). This plot is called a surface plot, and is used to identify clusters or patches of values that are similar. Clustering of similar values gives an indication of positive spatial autocorrelation, while values that change sharply within small local radii give an indication of negative spatial autocorrelation.
In addition to visualizing spatial autocorrelation, we can measure it formally using Moran's I.

2.2.2 Moran's I

A common way to measure spatial autocorrelation globally is to use Moran's I (Moran, 1950). Moran's I is an adaptation of the popular Pearson product moment correlation coefficient that allows us to measure spatial autocorrelation for a univariate series (Moran, 1950). The statistic is,

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.8}$$

where n is the number of observations, w_{ij} is the weight between observations i and j, and S_0 is the sum of all the weights. The choice of weight function between observations is important, and should take into account how close two observations are in space. By Tobler's first law (Tobler, 1970), observations that are closer in space are expected to have similar observed values, and are given a larger weight compared to observations that are further apart. A function of the inverse of the distance between observation locations is a popular choice for the weights. Other weight functions have been used in the literature, for example $w_{ij} = \exp(-\frac{d_{ij}}{d})$, which specifies quasi-global correlation between points derived from maximum entropy models (Chen, 2012).
Moran's I takes values between -1 and 1. If I is positive, then there is positive spatial autocorrelation within the sample, whereas, if it is negative,

then there is negative spatial autocorrelation. If I is equal to its expected value, then there is no spatial autocorrelation within the sample.
A test is defined for whether significant spatial autocorrelation exists by testing the null hypothesis that there is no spatial autocorrelation (ie. $I_{\text{obs}} = \text{E}(I)$), against an alternative hypothesis that there is (or the one-sided hypotheses that there is positive spatial autocorrelation or negative spatial autocorrelation). Standardizing I by its expected value and standard error, produces a convenient test statistic,

$$z_I = \frac{I - \text{E}(I)}{\sqrt{\text{Var}(I)}}, \tag{2.9}$$

where z_I is the test statistic with standard Normal distribution, $\text{E}(I)$ is the expected value of Moran's I, and $\sqrt{\text{Var}(I)}$ is the standard error. The expected value of Moran's I is defined as,

$$E(I) = \frac{-1}{n-1}, \tag{2.10}$$

where, n is the number of observations (Moran, 1950). The standard error for Moran's I is $\sqrt{E(I^2) - E(I)^2}$, and can be calculated analytically, or by bootstrapping.
The next section introduces tools for identifying non-stationarity in point referenced data.

2.3 Detecting Non-stationarity

2.3.1 Empirical covariance function

An empirical covariance function can be used to informally establish if an observed spatial process is stationary or non-stationary (Gelfand et al., 2010). The motivation behind this is that if the empirical covariance functions look different between different sub-regions of the data, then this provides evidence for non-stationarity. This involves assigning each pair

of observations to a distance class based on the distance between the locations of each pair of observations. An arbitrary number of classes, K, is chosen. The correlation between pairs of observations is plotted by the distance classes and any differences in the characteristics of the empirical covariance functions over distance can be observed for different classes. The empirical covariance function is evaluated for each distance class k by,

$$\hat{C}(\boldsymbol{h}_k) = \frac{1}{N_k} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(\boldsymbol{h}_k)} (Y(\boldsymbol{s}_i) - \bar{Y})(Y(\boldsymbol{s}_j) - \bar{Y}), \tag{2.11}$$

where $\boldsymbol{h}_k$ is the lag (distance between the locations of observations i and j in distance class k), N_k is the number of pairs of observations in distance class k, ($\boldsymbol{s}_i$ and $\boldsymbol{s}_j$) is a pair of locations for observations i and j in k, and Y and $\bar{Y}$ are the observed and average observations. By plotting the empirical covariance function against $\boldsymbol{h}_k$, we can get an informal perspective of the true covariance functions as distance between locations increase, across the whole study region. Non-stationarity is alluded to if the covariance functions look different between sub-regions.
A more formal way of identifying non-stationarity was proposed by Brunsdon et al. (1996), and we present it next.

2.3.2 Geographically weighted regression

The geographically weighted regression (GWR) model extends the traditional framework for regression by allowing local parameters to be estimated rather than global ones. The parameters are assumed to be functions of the locations on which the observations are obtained. The model for a simple regression is re-written as,

$$y_i = \sum_{k=0}^{p} \beta_{ik} x_{ik} + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{2.12}$$

where β_{ik} is the value of the kth parameter at location i, and x_{ik} is the value of kth covariate measured at location i. This model recognizes that spatial

variations in relationships might exist and provides an easy way to measure them. The calibration of Equation 2.12 implicitly assumes that data observed close to location i have more influence in the estimation of the β_{ik}'s than data located further from i. Essentially, the equation measures the relationships inherent in the model around each location i. Hence the parameters are estimated by weighted least squares. An observation is weighted in accordance with its proximity to location i, so that the weighting of an observation is no longer constant in the calibration but varies with i. That is,

$$\hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^T \boldsymbol{W}(i) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(i) \boldsymbol{Y}, \tag{2.13}$$

where $\boldsymbol{X}$ is an $n \times p + 1$ design matrix of covariates, and $\boldsymbol{W}(i)$ is a weight matrix, with zero as the off-diagonal elements and whose diagonal entries denote the geographical weighting of each of the n observed data for regression point i.

There are a range of choices for the spatial weight matrix $W(i)$. When the diagonal entries are all equal to 1, then we arrive at the ordinary least squares framework. A weight function that combats the discontinuities of weights is,

$$w_{ij} = \exp(-\theta d_{ij}^2), \tag{2.14}$$

where w_{ij} is the weight between a specific point in space j at which data are observed and any point in space i at which parameters are estimated. Here, d_{ij} is the distance between i and j, and $\sqrt{\frac{\theta}{2}}$ is called the bandwidth. If i and j both happen to be a point in space at which data are observed, the weighting at that point will be 1, and the weighting of other data will decrease according to a Gaussian curve as the distance between points increases.

We then calibrate the spatial weight matrix (ie. choose θ). The most convenient way to do so is to use generalized cross validation (Brunsdon et al.,

1996). This involves minimizing,

$$\text{GCV} = \frac{n}{(n - \text{tr}(\boldsymbol{H}))^2} \sum_{i=1}^{n} (y_i - \hat{y}(\theta))^2, \tag{2.15}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{W}(i) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(i)$, is the hat matrix, and $\hat{y}(\theta) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The variability of the local estimates can be used to examine the legitimacy of making a stationarity assumption. For a given independent variable k, at a given location i, suppose that $\hat{\beta}_{ik}$ is the GWR estimate of β_{ik}. If a value of this estimate is taken for each regression point (say n), then an estimate of variability in the parameter is given by the standard deviation of the n parameter estimates. By comparing this observed measure of variability to a distribution of variability measures under a null hypothesis of stationarity, a p-value can be obtained that describes the probability of observing such a variation in local parameter estimates from a stationary process. Algorithm 1 outlines the procedure used to calculate the p-values for the test.

Algorithm 1 Test for spatial non-stationarity

1: Find the optimal value for θ by minimizing Equation 2.15
2: Fit the GWR model using the optimal value of θ.
3: Calculate the observed standard deviation of each n estimates of each regression parameter.
4: Reshuffle the locations and assign them to the dependent and independent variables.
5: Refit the GWR model using the optimal value of θ found before, and calculate the standard deviations
6: Repeat (4) and (5) 1000 times to obtain a bootstrap sample of the null distribution.
7: Calculate the p-value by finding the proportion of bootstrapped standard deviations that are greater than the observe standard deviations.

We present the geostatistical model in the next section.

2.4 Geostatistical model

In Chapter's 3 and 4, we are interested in modelling the covariance structure of observed univariate spatial and spatio-temporal processes, $\{y(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$, and $\{y(\boldsymbol{s},t) : (\boldsymbol{s},t) \in \mathbb{R}^d \times \mathbb{R}\}$, respectively. A popular and well discussed kind of model, due to its flexibility in modelling the effect of relevant covariates as well as time and space dependence, is defined in several papers, in particular, Cameletti et al. (2011) and Sahu & Bakar (2012). The model is often referred to as a geostatistical model, and we used this term throughout the thesis. We defined the model separately for both spatial and spatio-temporal processes.

2.4.1 Geostatistical model for a spatial process

We assume that $y(\boldsymbol{s}_i)$, measured at location s_i where $i = 1, ..., n$, can be modelled by a Gaussian process, with measurement equation,

$$y(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \zeta(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i), \tag{2.16}$$

where $\mu(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)\boldsymbol{\beta}$ and $\boldsymbol{x}(\boldsymbol{s}_i) = (1, x_1(\boldsymbol{s}_i), ..., x_p(\boldsymbol{s}_i))$ denotes the $(p+1)$-dimensional vector of covariates for location $\boldsymbol{s}_i$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is the coefficient vector. Furthermore, the measurement error (nugget effect), $\varepsilon(\boldsymbol{s}_i)$, is modelled independently as a white noise process, $\mathrm{N}(0, \tau^2)$. Lastly, the spatial error, $\zeta(\boldsymbol{s}_i)$, is modelled by a zero-mean Gaussian distribution. It is characterized fully by the spatial covariance function,

$$\mathrm{Cov}(\zeta(\boldsymbol{s}_i), \zeta(\boldsymbol{s}_j)) = \sigma^2 R(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi}), \tag{2.17}$$

for $i \neq j$, and where σ^2 is the spatial variance parameter and $R(.)$ is a correlation function that depends on parameter vector $\boldsymbol{\phi}$, such that the resulting correlation matrix, R is positive definite. We make no assumption of spatial stationarity or isotropy, as evidenced by the spatial covariance function in Equation 2.17.

By collecting all the observations measured in a vector denoted by $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))'$, we can write,

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \tag{2.18}$$

where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{X} = (\boldsymbol{x}(\boldsymbol{s}_1)', ..., \boldsymbol{x}(\boldsymbol{s}_n)')'$, the measurement error follows $\boldsymbol{\varepsilon} \sim \text{MVN}(\boldsymbol{0}, \tau^2\boldsymbol{I}_n)$, the spatial process follows $\boldsymbol{\zeta} \sim \text{MVN}(\boldsymbol{0}, \sigma^2\boldsymbol{R})$, and $\boldsymbol{I}_n$ is the identity matrix with dimension n.
Furthermore, let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \tau^2, \boldsymbol{\phi}')'$ denote the vector of parameters. It is then implied that, from Equation 2.18, the marginal distribution of $\boldsymbol{y}$ (given the parameters), is,

$$\boldsymbol{y}|\boldsymbol{\theta} \sim \text{MVN}\Big(\boldsymbol{\mu}, \sigma^2\boldsymbol{R} + \tau^2\boldsymbol{I}_n\Big). \tag{2.19}$$

We used and modified Equation 2.18 in Chapter's 3 and 4 to model various sets of simulated and real data to analyze our proposed methodologies.

2.4.2 Geostatistical model for a spatio-temporal process

The geostatistical model for a spatial process can be extended to incorporate temporal dependence. We assume that $y(\boldsymbol{s}_i, t)$, measured at location s_i where $i = 1, ..., n$ and time $t = 1, ..., T$, can be modelled by a Gaussian process, with measurement equation,

$$y(\boldsymbol{s}_i, t) = \mu(\boldsymbol{s}_i, t) + Z(\boldsymbol{s}_i, t) + \varepsilon(\boldsymbol{s}_i, t), \tag{2.20}$$

where $\mu(\boldsymbol{s}_i, t) = \boldsymbol{x}(\boldsymbol{s}_i, t)\boldsymbol{\beta}$ and $\boldsymbol{x}(\boldsymbol{s}_i, t) = (1, x_1(\boldsymbol{s}_i, t), ..., x_p(\boldsymbol{s}_i, t))$ denotes the $(p+1)$-dimensional vector of covariates for location $\boldsymbol{s}_i$ at time t, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is the coefficient vector. Furthermore, the measurement error (nugget effect), $\varepsilon(\boldsymbol{s}_i, t)$, is modelled independently as a white noise process, $\text{N}(0, \tau^2)$. Lastly, $Z(\boldsymbol{s}_i, t)$ is a realization of a spatio-temporal process and is modelled by a Gaussian process that changes in time with first order autoregressive dynamics, and coefficient ρ, given by,

$$Z(\boldsymbol{s}_i, t) = \rho Z(\boldsymbol{s}_i, t-1) + \zeta(\boldsymbol{s}_i, t), \quad t = 2, ..., T. \tag{2.21}$$

where $|\rho| < 1$, and $Z(\boldsymbol{s}_i, 1)$ is such that,

$$Z(\boldsymbol{s}_i, 1) \sim \mathrm{N}\left(0, \sigma^2 \frac{R(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi})}{1-\rho^2}\right). \tag{2.22}$$

Further, $\zeta(\boldsymbol{s}_i, t)$ is modelled by a zero-mean Gaussian distribution, in which we assume temporal independence. It is characterized fully by the spatio-temporal covariance function,

$$\mathrm{Cov}(\zeta(\boldsymbol{s}_i, t), \zeta(\boldsymbol{s}_j, t^*)) = \begin{cases} 0 & \text{if } t \neq t^* \\ \sigma^2 R(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi}) & \text{if } t = t^*, \end{cases} \tag{2.23}$$

for $i \neq j$, and where σ^2 is the spatial variance parameter and $R(.)$ is a correlation function that depends on parameter vector $\boldsymbol{\phi}$, such that the resulting correlation matrix, R is positive definite. In this thesis, we implicitly make the assumption that the overall spatial covariance structure of the data process is constant over time. Also note that we make no assumption of spatial stationarity or isotropy, as evidenced by the spatial covariance function in Equation 2.23.

By collecting all the observations measured at time t in a vector denoted by $\boldsymbol{y}_t = (y(\boldsymbol{s}_1, t), ..., y(\boldsymbol{s}_n, t))'$, we can write

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{Z}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{MVN}(\boldsymbol{0}, \tau^2 \boldsymbol{I}_n), \tag{2.24}$$

for $t = 1, ..., T$, where $\boldsymbol{\mu}_t = \boldsymbol{X}_t\boldsymbol{\beta}$, $\boldsymbol{X}_t = (\boldsymbol{x}(\boldsymbol{s}_1, t)', ..., \boldsymbol{x}(\boldsymbol{s}_n, t)')'$, and $\boldsymbol{I}_n$ is the identity matrix with dimension n. As before, the spatio-temporal process is decomposed into spatial and temporal terms,

$$\boldsymbol{Z}_t = \rho \boldsymbol{Z}_{t-1}, + \boldsymbol{\zeta}_t, \quad \boldsymbol{\zeta}_t \sim \mathrm{MVN}(\boldsymbol{0}, \sigma^2 \boldsymbol{R}) \tag{2.25}$$

for $t = 1, ..., T$.

Also let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2_\varepsilon, \rho, \sigma^2, \tau^2, \boldsymbol{\phi})'$ denote the vector of parameters. It is then implied that

$$\boldsymbol{y}_t | \boldsymbol{Z}_t, \boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\mu}_t + \boldsymbol{Z}_t, \tau^2 \boldsymbol{I}_n), \tag{2.26}$$

for $t = 1, ..., T$, and

$$\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}, \boldsymbol{\theta} \sim \text{MVN}(\rho\boldsymbol{Z}_{t-1}, \sigma^2\boldsymbol{R}), \tag{2.27}$$

for $t = 2, ..., T$, and the $\boldsymbol{Z}_1$ comes from the stationary distribution of the AR(1) process,

$$\boldsymbol{Z}_1|\boldsymbol{\theta} \sim \text{MVN}\Big(\boldsymbol{0}, \frac{\sigma^2}{1-\rho^2}\boldsymbol{R}\Big). \tag{2.28}$$

It is then implied that, from Equations 2.26 – 2.28, the marginal distribution of $\boldsymbol{y}_t$ (given the parameters), is,

$$\boldsymbol{y}_t|\boldsymbol{\theta} \sim \text{MVN}\Big(\boldsymbol{\mu}_t, \frac{\sigma^2}{1-\rho^2}\boldsymbol{R} + \tau^2\boldsymbol{I}_n\Big). \tag{2.29}$$

We also used and modified Equation 2.24 in Chapter's 3 and 4 to model various sets of simulated and real data in order to analyze our proposed methodologies.

2.4.3 Covariance functions

There are many types of valid covariance functions, and they have been used in a variety of applications (Banerjee et al., 2014). One of the most common stationary covariance functions used to model spatial and spatio-temporal data is the Matèrn function,

$$\text{Cov}(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) = C(\boldsymbol{d}) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\psi\boldsymbol{d})^\nu K_\nu(\psi\boldsymbol{d}), \tag{2.30}$$

for $\psi > 0$ and $\nu > 0$, where $\Gamma(\nu)$ is the standard gamma function and K_ν is the modified Bessel function of the second kind with order ν (Matérn, 1960; Cressie, 1993). The parameter ψ controls smoothness of the rate of decay of the correlation as the distance between locations $\boldsymbol{d}$ increases and the parameter ν controls smoothness of the random field. When the smoothness parameter ν is set to specific values, closed form expressions can be obtained for the covariance function. For example, when $\nu = \frac{1}{2}$, the

Matèrn covariance function has closed form,

$$C(\boldsymbol{d}) = \sigma^2 \exp\left(- \frac{\boldsymbol{d}}{\psi} \right), \tag{2.31}$$

also known as the exponential covariance function. We used Equation 2.31 in various ways in Chapter's 3, 4, 5, and 6.

2.4.4 Cholesky factorization

In this thesis, we perform several simulation experiments to evaluate the viability of our proposed methodologies. In order to simulate a vector of spatial point reference observations, we use Cholesky factorization to draw from a Gaussian process, specifically the geostatistical models defined by Equations 2.18 and 2.24, with a specified mean vector and covariance matrix. Rue & Held (2005) provided simple algorithms for such computations, which have been built and implemented in various software, including R.
In general, we can decompose a matrix $\boldsymbol{V}$ into a lower triangular matrix and its transpose,

$$\boldsymbol{V} = \boldsymbol{L}\boldsymbol{L}^T, \tag{2.32}$$

where $\boldsymbol{L}$ is the lower triangular matrix. The lower triangular matrix retains the band structure from the orignal matrix, which allows computations to be carried out on $\boldsymbol{L}$. This has been shown to increase computational efficiency. We define the Cholesky factorization algorithm of Rue & Held (2005) below, that we used to sample from a Gaussian process with a mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
The following section is dedicated to Bayesian methods.

2.5 Bayesian methods

The geostatistical models described in Section 2.4 are examples of hierarchical models. The benefit of specifying a model with a hierarchical struc-

Algorithm 2 Sampling from a Gaussian process, $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

1: Compute the Cholesky factorization, $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$
2: Sample $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$
3: Solve $\boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{z}$
4: Compute $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{v}$
5: Return $\boldsymbol{x}$

ture is in the ability to explicitly incorporate different variance components of the response. For example, in Equation 2.18, we see that the variance of the dependent variable $\boldsymbol{y}$ is decomposed into two components, one for the spatial process, and one for the measurement process. Bayesian methods can be used to estimate hierarchical models and allow for easier computation of parameter estimates, compared to maximum likelihood estimation. Further, Bayesian hierarchical models have been used to estimate the parameters of geostatistical models in the literature (Gelfand & Banerjee, 2017). As such, we decided to use the Bayesian approach in this thesis.
We use the next section to lay the groundwork for the Bayesian approach.

2.5.1 Bayesian estimation

Let $\boldsymbol{y}$ be a vector of observations from some distribution depending on fixed potential predictors $\boldsymbol{x}$ and unknown parameters $\boldsymbol{\theta}$. We first start with a joint probability model for $\boldsymbol{y}$, $\boldsymbol{x}$, and $\boldsymbol{\theta}$, given by,

$$\pi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x}), \tag{2.33}$$

where we refer to $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ as the data likelihood and $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ as the prior distribution.
By simply conditioning the joint distribution of unknown parameters on the observed data $\boldsymbol{y}$, we arrive at an expression for the posterior density of the parameters:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \frac{\pi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{y}, \boldsymbol{x})} = \frac{f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{y}, \boldsymbol{x})}, \tag{2.34}$$

where $\pi(\boldsymbol{y}, \boldsymbol{x}) = \int \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$ is the marginal distribution of the data, which does not depend on any parameters. As such, we can rewrite the posterior density in its most recognizable form:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) \propto f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x}). \tag{2.35}$$

Equations 2.34 and 2.35 provide the theoretical framework for Bayesian statistics (Gelman et al., 2014). From here, we can build hierarchical models.

2.5.2 Bayesian hierarchical model

It is common practice to cast univariate geostatistical models as hierarchical models (Mukhopadhyay & Sahu, 2018; Gelfand & Banerjee, 2017; Banerjee et al., 2014; Cameletti et al., 2013; White & Ghosh, 2009). This lends itself nicely to the Bayesian methodology, which we adopt in this thesis. This offers the advantage for full and exact inference and proper assessment of uncertainty.

From equation 2.35, a Bayesian hierarchical model (BHM) for the geostatistical models in Section 2.4 can be built. Let $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T$ be a point referenced response variable where $y(\boldsymbol{s}_i)$ was observed at location $\boldsymbol{s}_i$. From Equation 2.18, we can obtain the likelihood function for $\boldsymbol{y}$, $f(\boldsymbol{y}|\boldsymbol{\zeta}, \boldsymbol{\theta}, \boldsymbol{x})$ by conditioning on the spatial process $\boldsymbol{\zeta}$, the parameter vector $\boldsymbol{\theta}$, and the predictors, $\boldsymbol{x}$. The value $\zeta(\boldsymbol{s}_i)$ is regarded as a random draw from a population distribution governed by some parameter vector $\boldsymbol{\theta}$, such that $\pi(\boldsymbol{\zeta}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(\zeta(\boldsymbol{s}_i)|\boldsymbol{\theta})$. Therefore, our BHM has the form:

$$\pi(\boldsymbol{\zeta}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) \propto f(\boldsymbol{y}|\boldsymbol{\zeta}, \boldsymbol{\theta}, \boldsymbol{x})\pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{x})\pi(\boldsymbol{\theta}|\boldsymbol{x}). \tag{2.36}$$

The posterior distributions for the parameters of a BHM are needed in order to perform Bayesian inference. From equation 2.36, three components are required to compute the posterior density.

2.5.3 Computational software

In this thesis we use the program `NIMBLE`, a system for building and sharing analysis methods for statistical models, especially for hierarchical models and computationally-intensive methods (NIMBLE Development Team, 2017). The program is built in `R`, but runs the models and algorithms using `C++` to increase computational efficiency. The benefit of using `NIMBLE` is that we can implement Markov chain Monte-Carlo (MCMC) sampling schemes to sample from the posterior distributions of the parameters. This is necessary because the full conditional distributions for each parameter are non-standard and do not have closed forms for estimation.

2.5.4 Model diagnostics

Trace plots

Trace plots display the parameter value at each iteration of sampling and are useful for visually assessing the convergence of the simulated draws to a posterior distribution. They allow us to check whether any values are rejected repeatedly causing the chain to become stuck on a single value (poor mixing). When poor mixing occurs, a particular value may be over-represented in the posterior sample.

In addition to monitoring chain mixing, trace plots allow us to check if any patterns are present. Clear patterns in a trace plot indicate that the algorithm may not have converged. In addition, if we use multiple chains and observe that they traverse different parts of the parameter space, then this is also indicative of non-convergence. To rectify these problems, we can increase the number of iterations.

Posterior density plots

Density plots of the posterior parameters are also useful for visualizing the quality of the posterior draws. The shapes of the densities are dependent on the distributions involved in the construction of the posteriors. However, if large and erratic peaks are observed in the density plots, then this indicates a lack of convergence to a single target distribution.

Autocorrelation plots

When consecutive values in a Markov chain are highly correlated, then traversing the sample space will be slow, leading to issues such as poor mixing. This is because the proposal parameters are more likely to be close to the current state. A plot of the autocorrelation for each parameter chain can be used to assess whether the correlation between successive draws will cause issues. If high correlation is detected, then thinning the sample (taking every kth draw) will make the values less dependent (Link & Eaton, 2012).

Potential scale reduction factor

The potential scale reduction factor (PSRF), $\hat{R}$, is used to monitor convergence of a posterior distribution for a parameter, θ, to a stationary distribution. It is an estimate of the factor, by which the scale of the current distribution for a parameter might be reduced if simulations were continued in the limit $n \to \infty$.
To compute $\hat{R}$ for a posterior simulation of θ, we must first estimate the marginal posterior variance, Var$(\theta|y)$. This can be done by a weighted average of within-chain variance, and between-chain variance. Let θ_{ij} represent the posterior draw for iteration $i = 1, ..., n$ from chain $j = 1, ..., m$. Then the between-chain variance is defined as,

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2, \tag{2.37}$$

where,

$$\bar{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \text{and} \quad \bar{\theta}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_{.j}, \tag{2.38}$$

and the within-chain variance is defined as,

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \tag{2.39}$$

where,

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2. \tag{2.40}$$

The marginal posterior variance is given by,

$$\widehat{\text{Var}}(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{2.41}$$

Then, the PSRF is given by,

$$\hat{R} = \sqrt{\frac{\widehat{\text{Var}(\theta|y)}}{W}}. \tag{2.42}$$

This expression converges toward 1 as n tends to ∞. As a result, if a parameter chain has a PSRF of 1 (or close to 1), then it can be assumed that the chain has converged to the target distribution.

2.5.5 Posterior prediction

We use the posterior predictive distribution to obtain a distribution of fitted values, $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_n)^T$. The posterior predictive distribution is given by,

$$f(\hat{\boldsymbol{y}}|\boldsymbol{y}) = \int f(\hat{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{2.43}$$

where $f(\hat{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\theta})$ is a normal distribution arising from the joint multivariate normal distribution of $\hat{\boldsymbol{y}}$ and the original data $\boldsymbol{y}$. We can readily obtain estimates of the posterior predictive distribution. Suppose we draw a posterior draw $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(L)}$ from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$. Then

we use composition sampling to draw $\hat{\boldsymbol{y}}$, one for each $\boldsymbol{\theta}^{(l)}$, that is, $\hat{\boldsymbol{y}}^{(l)} \sim f(\hat{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\theta}^{(l)})$. The resulting collection $\{\hat{\boldsymbol{y}}^{(1)}, \hat{\boldsymbol{y}}^{(2)}, ..., \hat{\boldsymbol{y}}^{(L)}\}$ is a sample from the posterior predictive density (Gelfand & Banerjee, 2017).

2.6 Model comparison and assessment

2.6.1 Predictive accuracy

In this thesis, we placed particular emphasis on assessing a (Bayesian) models ability to make accurate predictions. In this section, we introduce the concept of predictive accuracy. We then list several measures of predictive accuracy that we use to compare the models fitted in this thesis.
Once a model has been fit, it is necessary to measure the model's predictive accuracy. Predictive accuracy allows for assessment of a model's goodness of fit, and can be used in the process of model comparison and selection Vehtari & Gelman (2014). We can measure a model's predictive accuracy in different ways, which are each tailored toward the model application. In this thesis, we measure predictive accuracy using two methods.
The first measure that we calculate is the root mean square error (RMSE). Under the Bayesian context, we compute the RMSE for a particular draw from the posterior distribution, $\boldsymbol{\theta}^{(l)}$, as

$$\text{RMSE}^{(l)} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(l)} \right)^2 \right)^{\frac{1}{2}}, \tag{2.44}$$

where $\hat{\boldsymbol{y}}^{(l)} = (\hat{y}_1^{(l)}, ..., \hat{y}_n^{(l)})^T$ for $l = 1, ..., L$.
In addititon to RMSE, we compute the mean absolute error (MAE), a more natural measure of average error magnitude (Willmott & Matsuura, 2005). For a particular draw from the posterior distribution, $\boldsymbol{\theta}^{(l)}$

$$\text{MAE}^{(l)} = \frac{1}{n} \sum_{i=1}^{n} \text{abs}\big(y_i - \hat{y}_i(\boldsymbol{\theta}^{(l)})\big). \tag{2.45}$$

2.6.2 Average estimation error of the covariance matrix

The covariance matrix plays a central role in this thesis. In addition to predictive accuracy, we also evaluate a models ability to accurately estimate the covariance matrix. The average estimation error of the covariance matrix is given by,

$$\text{COV} = \log\left(\frac{1}{L}\sum_{l=1}^{L} n^{-\frac{1}{2}}||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}^{(l)}) - \boldsymbol{\Sigma}||_{\text{F}}\right), \tag{2.46}$$

(Lan et al., 2018).

2.6.3 Residual spatial autocorrelation

The last measure that we used to evaluate a proposed model is Moran's I on the residuals. By calculating Moran's I on the residuals, we obtain a measure of the amount of spatial autocorrelation that is not accounted for by the model. Moran's I on the residuals is calculated in the same way as in Section 2.2.2, except that the observed values are replace by the posterior distributions of each residual. That is,

$$I^{(l)} = \frac{n}{S_0}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(l)}(r_i^{(l)} - \bar{r^{(l)}})(r_j^{(l)} - \bar{r^{(l)}})}{\sum_{i=1}^{n}(r_i^{(l)} - \bar{r^{(l)}})^2}, \tag{2.47}$$

where $I^{(l)}$ is the value of Moran's I on the residuals calculated from the lth draw of the posterior residuals,

$$r_i^{(l)} = y_i - \hat{y}_i^{(l)}. \tag{2.48}$$

We now introduce the two datasets that we apply our methods to throughout this thesis.

2.7 Datasets

2.7.1 New Zealand particulate matter

Air pollution has been shown to have negative effects on human health, including premature mortality, as well as lung and heart problems. A review also found that particulate matter (PM10 and PM2.5, Figure 2.1) causes lung cancer (Loomis et al., 2013).

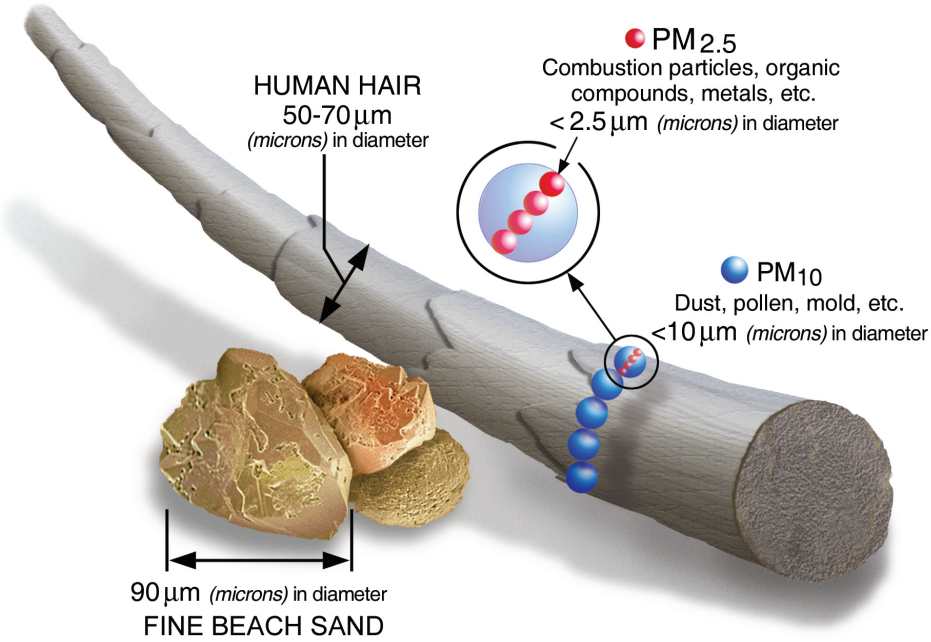


Figure 2.1: Illustration that puts the size of PM10 and fine particulate matter, PM2.5, into perspective. The image was obtained from US Environmental Protection Agency (2018).

In 2012, air pollution from human made PM10 in New Zealand was associated with an estimated 1000 premature deaths, 520 extra hospital admissions for cardiovascular and respiratory diseases, and 1.35 million restricted activity days (when symptoms would prevent an individual from

performing their usual daily activities) (Ministry for the Environment & Statistics New Zealand, 2015). The greatest contributor to human-made PM10 in New Zealand is burning wood and coal for home heating, with an annual contribution of 57.5%, rising to 79.2% in winter (Ministry for the Environment & Statistics New Zealand, 2015). New Zealand's 'resource management regulations' (Ministry for the Environment & Statistics New Zealand, 2007) state that the concentration of PM10 cannot surpass 50 micrograms per metre cubed (expressed as a 24 hour mean). Currently, New Zealand does not have any standards on annual mean PM10, but the World Health Organization's (WHO, 2006) guidline is 20 micrograms per metre cubed. PM10 concentrations above this threshold are considered harmful to human health. Under the resource management act (1991), regional councils and unitary authorities are responsible for managing air quality. Subsequently, monitoring stations have been placed at 40 fixed locations throughout New Zealand where pollution levels are highest. Annual average PM10 concentrations from these monitoring stations were obtained from the Ministry for the Environment data service (Ministry for the Environment, 2015) for the year 2013.

Figure 2.2 shows the distribution of the monitoring stations. In addition, the colour gradient displays the concentration of PM10, with blue corresponding to a lower annual concentration in 2013 and red corresponding to a higher annual concentration in 2013. We can see that monitoring stations in the South Island of New Zealand recorded higher concentrations of PM10 than stations in the North Island. The lowest recorded annual concentration of PM10 was observed at a monitoring station in the Wellington region.

The level of ambient PM10 depends on how much pollution is being produced as well as other factors like the weather and local geography. For example, windy conditions help to move pollution away. However, features such as valleys can cause pollution to linger. Temperature inversions that occur during cold, calm conditions can trap in pollutants, causing

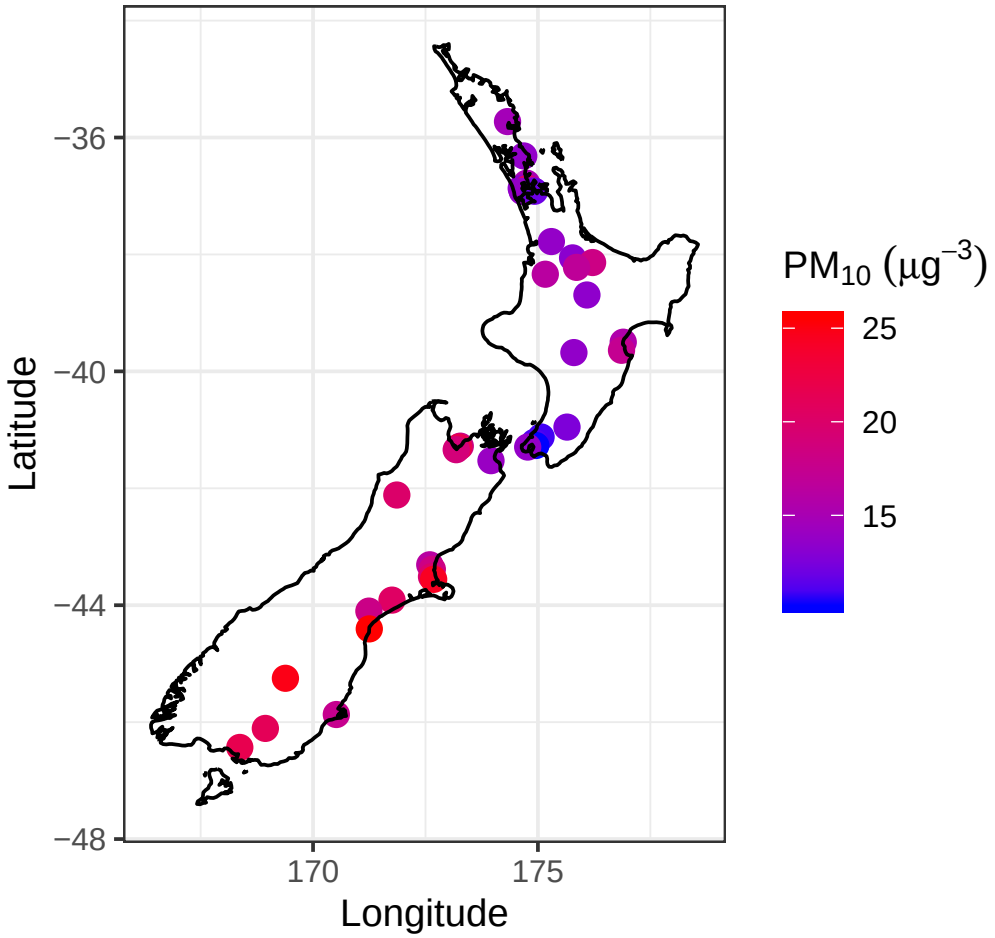


Figure 2.2: Annual average particulate matter (PM10) concentration for 2013 observed at monitoring stations in New Zealand. Red represents higher levels of PM10 concentration and blue represents lower levels of PM10 concentration.

levels to be high. Years where there have been a lot of temperature inversions often have worse air quality, and produced more PM10 in colder years. To account for covariate effects, monthly average temperature (in °C) and wind speed (m/s) measurements were obtained from the NIWA Climate Database. These values were measured at 2089 stations throughout New Zealand. Monthly averages were obtained from these locations and converted into annual averages inline with the temporal scale of the PM10 measurements. The covariates were available at different locations from the PM10 values, the so-called "change of support problem" (Banerjee et al., 2014). An estimate of the covariate value was obtained at each PM10 location by first creating a grid of $1^\circ \times 1^\circ$ (approximately 111 km $\times$ 111 km) cells over New Zealand. The covariate value assigned to each PM10 monitoring station was then the average of all values in the same grid cell as the monitoring station.

In this thesis, we fitted several models to the 2013 PM10 point referenced data to illustrate our proposed methodologies in Chapters 3, 4, and 5. The PM10 data was suitable to test our methods because the data exhibited features like significant spatial autocorrelation and non-stationarity. Evidence of spatial autocorrelation can be gleaned from Figure 2.2, where clusters of higher values are seen in the south and clusters of lower values can be seen in the north. Furthermore, Moran's I confirms the presence of significant spatial autocorrelation. Moran's I was calculated to be $M = 0.3577$, with a corresponding p-value for the two-sided test of significant spatial autocorrelation of 3.23×10^{-8}. In addition to spatial autocorrelation, there is evidence for non-stationarity in the PM10 data. We found evidence for non-stationarity when we performed the test for non-stationarity that was described in Section 2.3.2. We calculated a p-value of $p = 0.03$, suggesting significant non-stationarity.

2.7.2 Sub-Antarctic hoki

Research trawl surveys of the sub-Antarctic region were carried out by the National Institute of Water and Atmospheric Research (NIWA) for the Ministry of Primary Industries, New Zealand (MPI). The dataset is a time series that has been accumulated from the summers of 1991 to 1993, and then again from 2000 to 2008 (Bagley et al., 2013). It contains point referenced data on catch weight for multiple species in the sub-Antarctic. The purpose of the surveys was primarily to estimate abundance of a particular fish species, Macruronus novaezelandiae, commonly known as hoki. For this reason, we focused on the positive catch weight data for hoki in this thesis.

The data were stratified using a 2-phase adaptive survey method proposed in Francis (1984), with the intention to reduce variation in biomass estimates. Figure 1.4 illustrates the stratification and shows the observed hoki catch weight in kilograms and location for each of the 1255 trawls. In addition, the year of the trawl is indicated by the colour, with red representing the earlier years, and pink indicating more recent trawls. From the graphic, we can glance that trawls near Puysegur Bank measured the largest catch weights for hoki.

For our applications, we focused on the trawls that occurred between 2000 to 2008, in order to have a consistent annual time series. We fitted several models to the data to illustrate our proposed methodologies in Chapters 4, 5, and 6. For Chapters 4 and 6, the models required that measurements be observed at the same locations across time. The hoki dataset does not meet this criteria. In order to obtain repeated measurements at the same locations annually, catch weight locations within a stratum were gridded. The strata were gridded in such a way that within each grid there was at least one catch weight observation per year. The median longitude and latitude of all observations within a grid was taken as the grid center, g. The 38 grid centers are shown in Figure 2.3.

It should be noted that not all strata were used in the grid construction. For

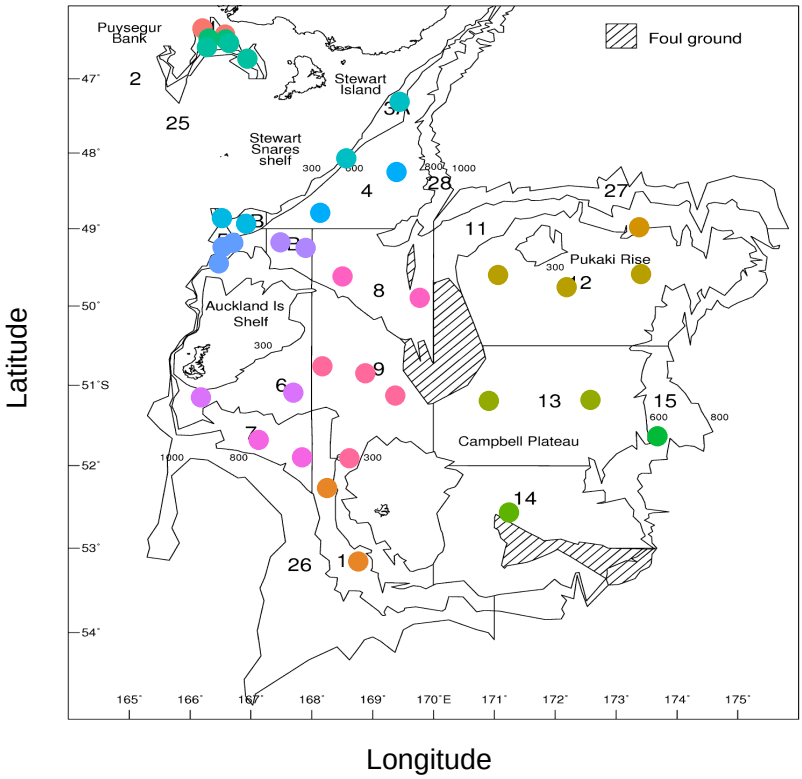


Figure 2.3: Map of 38 grid centers that summarise the trawls where hoki was caught in the sub-Antarctic region.

strata 25 – 28, there were years that trawls did not occur, and were therefore excluded from the final dataset. The weighted mean of hoki catch weight observations within a grid was assigned as the catch weight for the entire grid. The mean was weighted by distance from the grid center, with catch weight observations closer to the grid center given more weight. A weighted mean is used to allow observations located closer to the grid center to contribute more to the grid mean than those further away. In addition, the weighted mean depth of each trawl within a grid was assigned as the depth for the entire grid in the same fashion.

Once again, the hoki data was suitable to test our methods because the data exhibited significant spatial autocorrelation and non-stationarity. We produced interpolated surface plots of the hoki catch weight (in log scale) for each year and these are displayed in Figure 2.4. From the plots, we observed evidence for spatial autocorrelation within each year in the form of clusters of high and low catch weights across the sub-Antarctic region. Furthermore, Moran's I confirmed the presence of significant spatial autocorrelation for the years 2000, and 2003 – 2006 (see Table 2.1). Non-stationarity was also detected for the years 2000 – 2005, and 2007, confirmed by the p-values for the test of non-stationarity, given in Table 2.1.

We now move on to the main Chapters of this thesis, in which we introduce new methodologies.

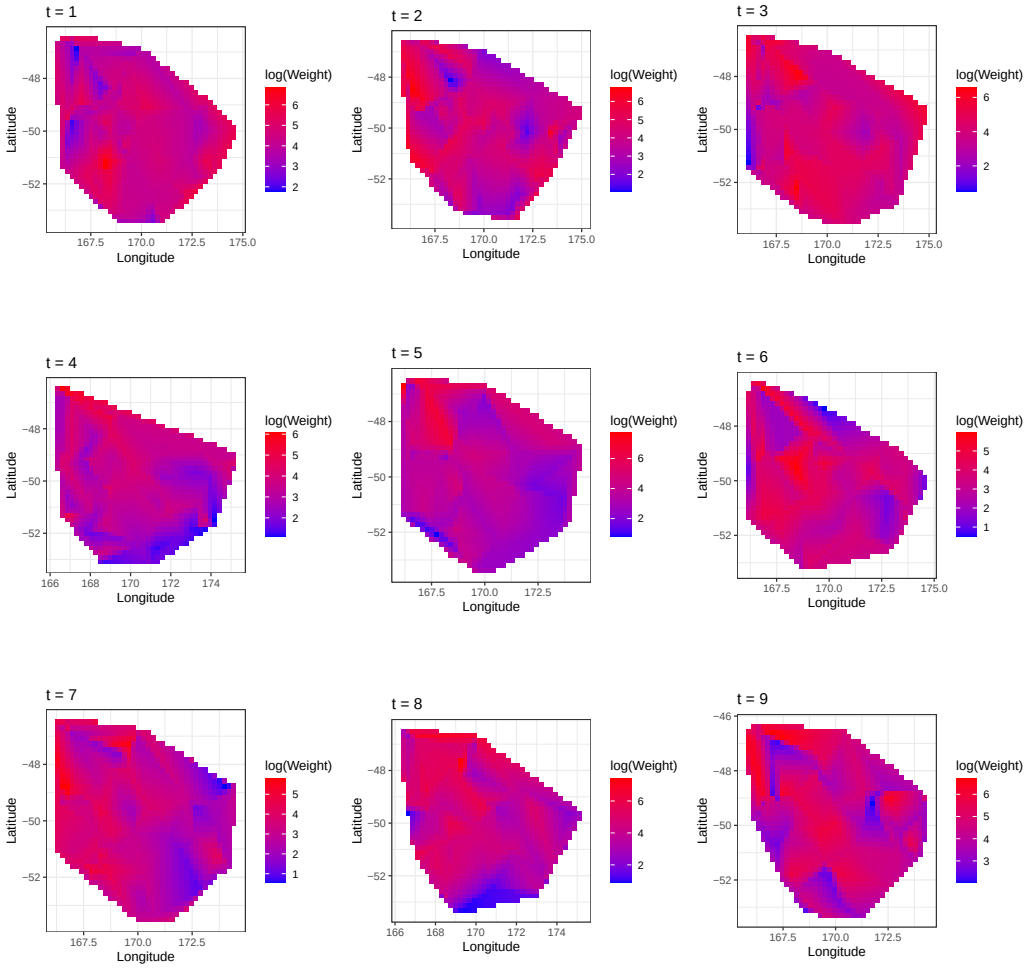


Figure 2.4: Surface plots for Akima interpolated hoki catch weight across the sub-Antarctic region for each year.

Table 2.1: Statistics and p-values for the tests of spatial autocorrelation and non-stationarity for each year of the hoki catch weight data.

	Spatial Autocorrelation		Non-stationarity
Year	Moran's I	p-value	p-value
2000	0.0721	1.208×10^{-5}	0.04
2001	0.00858	0.363	0.02
2002	0.00164	0.541	0.09
2003	0.197	$< 2.2 \times 10^{-16}$	0.025
2004	0.0679	1.298×10^{-8}	0
2005	0.0893	2.463×10^{-4}	0.07
2006	0.108	8.527×10^{-7}	0.125
2007	-0.00302	0.444	0.065
2008	-0.00374	0.726	0.195

Chapter 3

Partitioned geostatistical models for spatial and spatio-temporal data

Spatial and spatio-temporal phenomena are commonly modelled as Gaussian processes. One such construction is the geostatistical model defined by Equation 2.18 in Section 2. The geostatistical model has the benefit of modelling the spatial dependence structure using covariance functions. The most commonly used covariance functions impose an assumption of spatial stationarity on the process. That means the covariance between observations at particular locations depends only on the distance between the locations (Banerjee et al., 2014). It has been widely recognized that most processes manifest a spatially non-stationary covariance structure (Sampson, 2014). If the study domain is small in area or there is not enough data to justify more complicated non-stationary approaches, then stationarity may be assumed for the sake of mathematical convenience (Fouedjio, 2017). However, relationships between variables can vary significantly over space, and a 'global' estimate of the relationships may obscure interesting geographical phenomena (Brunsdon et al., 1996; Fouedjio, 2017; Sampson & Guttorp, 1992). In Section 1.2 we described why it

is important to carefully consider non-stationarity when estimating and making predictions from a modelled spatial or spatio-temporal process.

Literature on the geostatistical model for spatial and spatio-temporal processes suggests that non-stationarity can be considered in two ways. One approach involves using a covariance function that explicitly accounts for non-stationarity. This approach was investigated in (Paciorek & Schervish, 2006; Paciorek et al., 2013). The other approach uses non-parametric methods. A range of techniques have been described in the literature, such as partitioning, kernel smoothing, process convolution, and spatial deformation Sampson (2014); Fouedjio (2017). In this chapter, we focused on partitioning the spatial domain into heterogeneous sub-regions, within which stationarity is assumed.

We begin this chapter with a review on the literature around partitioning spatial and spatio-temporal data. We highlight the successes and limitations of several partitioning strategies applied to spatial and spatio-temporal data. In Section 3.2 we formally introduced the partitioned geostatistical model of Heaton et al. (2017) for spatial data. We contributed an extension of the partitioned geostatistical model to the spatio-temporal case. This is followed by Section 3.3, where we contributed a partitioning strategy that uses the K-means algorithm to partition the locations based on geographic features such as Euclidean distance. In Section 3.4.1 we outlined a simulation study that compared K-means partitioned geostatistical models fitted to two simulated spatial datasets, one generated with a stationary covariance structure, and the other generated with a non-stationary structure. The simulation was repeated for the spatio-temporal case in Section 3.4.2. A case study was carried out in Section 3.5. We fitted the proposed K-means partitioned geostatistical models to New Zealand particulate matter data. We concluded the chapter with comments and reflections in Section 3.6.

3.1 Literature review

One of the most widely used assumption in spatial and spatio-temporal statistics is that of (second-order) stationarity. In Chapter 2, we introduced the concept of stationarity in the context of Gaussian processes. The assumption of stationarity implies that the data are representative of a complete sampling of a single realization (Cressie, 1993). In other words, the mean of a stationary spatial or spatio-temporal process is constant (over space) and the covariance between elements of a dependent variable is a function only on their spatial separation. The assumption of stationarity is a popular one to make because a range of simple and easily interpretable covariance functions are available (Banerjee et al., 2014).

In reality, however, stationarity might not always be appropriately assumed. There are situations when stationarity can be assumed because the study domain is too small, the amount of data is too small to justify more complex models, or there are no other suitable alternatives (Fouedjio, 2017). However, covariance between observations over space and/or time are likely to change across space. While stationarity has its advantages, the limitations to the assumption are not ignorable. Using stationary modelling approaches when local influences or localized effects, such as topographic features, exist within the study domain can lead to less accurate prediction and an incorrect assessment of the prediction error (Fouedjio, 2017).

It has been recognized that most, if not all, spatial and spatio-temporal processes manifest spatially non-stationary covariance structure (Sampson, 2014). This is so, because these processes depend on underlying latent processes that change over space, such as topographic structure or geographic features (Fouedjio, 2017). As a result, the covariance of the observed process will be different depending on location. Fitting a global covariance function to model the spatial or spatio-temporal process that assumes stationarity would therefore be inappropriate.

One approach to overcome the issue of non-stationarity is through the method of partitioning (Fouedjio, 2017). This involves partitioning the spatial domain into sub-regions using an appropriate partitioning strategy. The motivation for partitioning is two-fold. Firstly, the size of the sub-regions are smaller than the entire study region, and become small enough to assume that the process is stationary within a region. Secondly, using an appropriate partitioning strategy would allow conditioning the observed process on an appropriate geographic feature or proxy. This would allow to account for a possible cause of the non-stationarity (Fouedjio, 2017). For both reasons, fitting region specific stationary covariance functions would become valid.

Partitioning of spatial processes has been examined and implemented in geostatistical literature. One strategy for partitioning involves forming sub-regions along natural boundaries where the covariance between observations is thought to, or known to change. An example of this type of partitioning is given in McBratney et al. (1991). It was stated that when mapping soil attributes, it is important to consider spatial heterogeneity across the study region. As such, they proposed partitioning the study area, where the content of soil surface was measured, into two regions based on topography. They fitted intrinsic random functions of order k within each region, and produced a predicted map of the soil attribute using kriging.

Another example is given in Atkinson & LLoyd (2007). They applied a partitioning model to elevation data from the Ordnance Survey Great Britain region ST92SW, to the east of Shaftesbury. The region of interest was partitioned via a traditional image-processing algorithm. The approach employed was a centroid-linkage region growing partitioning algorithm (Haralick & Shapiro, 1985). Different covariance structures were allowed to be fitted to each sub-region. The aim was to show how the covariance structure changes over the entire study region.

Gosoniu et al. (2009) modelled spatial heterogeneity in malaria prevalence

data in West Africa using a geostatistical model. In order to account for any changes in covariance across the study region, they used a fixed Bayesian partitioning method based on agro-ecological zones. The study region was split into four sub-regions with each sub-region corresponding to a different ecological zone. Stationary exponential covariance functions were fitted within each sub-region. Furthermore, the spatial covariance parameters varied by region.

These partitioning strategies, while easy to implement, could be considered subjective in that it is up to the researcher to make a decision on how to partition the observations. A data driven method was presented in Kim et al. (2005). Kim et al. (2005) proposed a Bayesian partition model that accounted for sharp transitions in covariance structure. The study region was partitioned using Voronoi tessellation (Green & Sibson, 1978). This was done so that within each sub-region the covariance structure could be assumed stationary. Between sub-regions, independence was assumed. In order to smooth the predicted spatial process at region boundaries, Bayesian model averaging was implemented. Kim et al. (2005) performed a range of simulation experiments to test their partitioning methodology, in which they found that partitioning by Voronoi tessellation allowed their model to infer the correct number of partitions used to generate the data. They also found that by partitioning, they avoided the computational issue of inverting large covariance matrices. In addition, they applied their model to soil permeability of the Schneider Buda oil field in Wood County, Texas, United States and found sufficient evidence against a model with no partition structure.

Another example was presented in Heaton et al. (2017). (Heaton et al., 2017) employed a spatially clustered Gaussian process model, in which the study region was partitioned into disjoint sets through a dissimilarity measure,

$$d_{ij} = \frac{|y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j)|}{||\boldsymbol{s}_i - \boldsymbol{s}_j||}, \tag{3.1}$$

where $y(\boldsymbol{s}_i)$ is a response observed at location $\boldsymbol{s}_i$, and $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ is the

Euclidean distance between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. This dissimilarity metric is motivated by spatial finite differences, which is an estimate of the directional derivative as described in Banerjee et al. (2003a) and Banerjee & Gelfand (2006). This dissimilarity metric tends to cluster observations based on the change in the spatial surface. In other words, the dissimilarity d_{ij} will be large when the covariance structure changes rapidly in the $\boldsymbol{s}_j - \boldsymbol{s}_i$ direction from $\boldsymbol{s}_i$, leading to $y(\boldsymbol{s}_i)$ and $y(\boldsymbol{s}_i)$ being assigned to different clusters. Cluster boundaries will be placed along directions of a large derivative. These large observed rates of change are natural regional boundaries because they represent areas where assumptions of isotropy and stationarity may not hold. However, this led to unclear geographic partitioning, where there was no way to clearly separate the partitions geographically. According to Heaton et al. (2017), the partitions should theoretically be stationary.

We proposed a simple method for modelling spatial processes without assuming global stationarity using partitioning. Partitioning offers flexibility in that we can reduce the largest distance between locations of observations, as well as easily incorporate geographic features or proxies into the partitioning function. Stationarity of the covariance structure can then be assumed for each sub-region because the area will be smaller, and by taking account of the topography through a covariate, removes that as factor causing non-stationarity. The partitioning method we proposed involves partitioning the observed process using K-means clustering, then fitting simple covariance function models to the observations in each partition. Furthermore, we explored the possibility that the covariance function need not be the same for all sub-regions. In this thesis we assumed conditional independence between sub-regions, however, we believe this assumption may be relaxed.

3.2 Partitioned geostatistical model

In this section, we introduced the partitioned covariance model framework of Heaton et al. (2017) for spatial processes. We then proposed an extension of this framework to the spatio-temporal case. In both cases, the framework assumes that the processes are Gaussian. We begin with the partitioned covariance model for spatial processes.

Heaton et al. (2017) presented a partitioned form of the geostatistical model described by Equation 2.18. Let $(y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))'$ be a vector of point referenced response variables observed at locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ within spatial domain $\mathcal{S} \in \mathbb{R}^2$. Then let $\{\boldsymbol{s}_i\}_{i=1}^N$ be the collection of locations and partition them into K distinct sub-regions, $\mathcal{S}_1, ..., \mathcal{S}_K$, such that $\bigcup_{k=1}^K \mathcal{S}_k = \mathcal{S}$, and $\mathcal{S}_{k_1} \bigcap \mathcal{S}_{k_2} = \varnothing$ for all $k_1 \neq k_2$, according to an appropriate partitioning strategy. We denote $\boldsymbol{y}_k = \{y(\boldsymbol{s}_i \in \mathcal{S}_k)\}$ as the vector of observations that belong to region $\mathcal{S}_k$ and denote n_k as the number of observations in region k, where $n = \sum_{k=1}^K n_k$. Then, by assuming conditional independence between the regions, we have for $k = 1, ..., K$,

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k, \tag{3.2}$$

where $\boldsymbol{X}_k = (\mathbf{1}_{n_k}, \boldsymbol{x}_{1k}, ..., \boldsymbol{x}_{Pk})$ is the $n_k \times (P+1)$ design matrix of covariates, $\boldsymbol{x}_{pk}$, with corresponding, sub-region specific coefficients, $\boldsymbol{\beta}_k = (\beta_{0k}, ..., \beta_{Pk})'$, $\boldsymbol{\zeta}_k$ is the spatial error term assumed to follow,

$$\boldsymbol{\zeta}_k \sim \mathrm{N}(\mathbf{0}, \sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k)), \tag{3.3}$$

with sub-region specific covariance matrix depending on covariance parameters $\boldsymbol{\phi}_k$, and $\boldsymbol{\varepsilon}_k$ is the measurement error term, assumed to follow $\boldsymbol{\varepsilon}_k \sim \mathrm{N}(\mathbf{0}, \tau_k^2\boldsymbol{I}_{n_k})$, with sub-region specific nugget variance τ_k^2. By collecting the observations from each sub-region, $\boldsymbol{y} = (\boldsymbol{y}_1', ..., \boldsymbol{y}_K')'$, we write from Equation 3.2,

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \mathbf{T}), \tag{3.4}$$

where $\boldsymbol{X}$ is a block-diagonal design matrix with $\boldsymbol{X}_k$ on the main diagonal, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_K')'$, and $\boldsymbol{\Sigma}$ and T are block-diagonal with main diagonal matrices $\sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k)$ and $\tau_k^2 \boldsymbol{I}_{n_k}$, respectively.

We note that in many applications, the coefficients $\{\boldsymbol{\beta}\}_{k=1}^K$ are assumed equal across regions, which represent the global effect of covariates on the response. However, assuming common coefficients rules out the possibility of local rather than global covariate effects.

We extended the partitioning framework of Heaton et al. (2017) to the spatio-temporal case. We assumed that temporal autocorrelation was satisfactorily accounted for by a first-order autoregressive process, such was the case in Equations 2.24 to 2.25. In this thesis, we made the assumption of no space-time interaction, and we focused on partitioning only over space.

Let $(y_t(\boldsymbol{s}_1), ..., y_t(\boldsymbol{s}_n))'$ be a vector of point referenced response variables observed at locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ at discrete time points $t = 1, ..., T$, within spatial domain $\mathcal{S} \in \mathbb{R}^2$, and within temporal domain $\mathcal{T} \in \mathbb{R}$. Then let $\{\boldsymbol{s}_i\}_{i=1}^n$ be the collection of locations and partition them into K distinct sub-regions, $\mathcal{S}_1, ..., \mathcal{S}_K$, such that $\bigcup_{k=1}^K \mathcal{S}_k = \mathcal{S}$, and $\mathcal{S}_{k_1} \bigcap \mathcal{S}_{k_2} = \varnothing$ for all $k_1 \neq k_2$, according to an appropriate partitioning strategy. We denote $\boldsymbol{y}_{kt} = \{y_t(\boldsymbol{s}_i \in \mathcal{S}_k)\}$ as the vector of observations that belong to region $\mathcal{S}_k$ and denote n_k as the number of observations in region k, where $n = \sum_{k=1}^K n_k$. Then, by assuming conditional independence between the regions, we have for $k = 1, ..., K$,

$$\boldsymbol{y}_{kt} = \boldsymbol{X}_{kt}\boldsymbol{\beta}_k + \boldsymbol{Z}_{kt} + \boldsymbol{\varepsilon}_{kt}, \tag{3.5}$$

where $\boldsymbol{X}_{kt} = (\boldsymbol{1}_{n_k}, \boldsymbol{x}_{1kt}, ..., \boldsymbol{x}_{Pkt})$ is the $n_k \times (P+1)$ design matrix of covariates, $\boldsymbol{x}_{pkt}$, at time t, with corresponding, sub-region specific coefficients, $\boldsymbol{\beta}_k = (\beta_{0k}, ..., \beta_{Pk})'$, $\boldsymbol{Z}_{kt}$ is the spatio-temporal process, with first-order autoregressive dynamics, $\boldsymbol{Z}_{kt} = \rho \boldsymbol{Z}_{kt-1} + \boldsymbol{\zeta}_k$, with $Z_{k1} \sim \mathrm{N}(\boldsymbol{0}, \frac{\sigma_k^2}{1-\rho^2}\boldsymbol{R}(\boldsymbol{\phi}_k))$, and $\boldsymbol{\zeta}_{kt} \sim \mathrm{N}(\boldsymbol{0}, \sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k))$, with sub-region specific covariance matrix depending on covariance parameters $\boldsymbol{\phi}_k$, and $\boldsymbol{\varepsilon}_k$ is the measurement error

term, assumed to follow $\boldsymbol{\varepsilon}_k \sim \mathrm{N}(\mathbf{0}, \tau_k^2 \boldsymbol{I}_{n_k})$, with sub-region specific nugget variance τ_k^2. By collecting the observations from each sub-region, and each time point, $\boldsymbol{y} = (\boldsymbol{y}'_{11}, ..., \boldsymbol{y}'_{K1}, ..., \boldsymbol{y}'_{KT})'$, we write from Equation 3.5,

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \mathrm{T}), \tag{3.6}$$

where $\boldsymbol{X}$ is a block-diagonal design matrix with $\boldsymbol{X}_k$ on the main diagonal, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, ..., \boldsymbol{\beta}'_K)'$, and $\boldsymbol{\Sigma}$ and T are block-diagonal with main diagonal matrices $\sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k)$ and $\tau_k^2 \boldsymbol{I}_{n_k}$, respectively.

We now present our proposed strategy for partitioning the study region.

3.3 Partitioning using the K-means algorithm

We assumed that the covariance structure is governed by location only or other geographical covariates that have spatial distributions. We used the K-means clustering algorithm to create sub-regions based on location or geographic covariates. By incorporating geographic covariates, we are able to account for non-stationarity explicitly.

K-means clustering is considered one of the simplest and most popular partitioning algorithms (Jain, 2010). The K-means algorithm was independently developed in different scientific fields by Steinhaus (1956), Lloyd (1982), Ball & Hall (1965), and MacQueen et al. (1967). As such, it has a rich and diverse history. Although K-means clustering was first proposed over 60 years ago, it is still one of the most widely used algorithms for clustering (Jain, 2010), due to its ease of implementation, simplicity, efficiency, and empirical success.

K-means clustering is an unsupervised learning algorithm, since there is no training or test set of observations to check whether the clustering assignments were correct. The number of partitions can be chosen arbitrarily based on whether or not you know how many clusters there should be, or it can be chosen using an explorative approach.

The K-means algorithm is presented in Algorithm 3. Let $\boldsymbol{\delta}_i$ be an appropriate feature (e.g. longitude, latitude) vector associated with the observation at location $\boldsymbol{s}_i$. Then let $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K$ be the centroids, where $\boldsymbol{\gamma}_k$ is the centroid of locations closest to cluster k. For partitioning a spatial or spatio-temporal process, we propose setting $\boldsymbol{\delta}_i = \boldsymbol{s}_i$.

Algorithm 3 K-Means Algorithm

1: Randomly select initial partition centroids, $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K \in \mathbb{R}^d$, where $\boldsymbol{\gamma}_k$ is a d-dimensional centroid vector for partition k..
2: Repeat until cluster assignments no longer change:

- for $i = 1, ..., N$, set

$$c_i := \arg\min_k (||\boldsymbol{\delta}_i - \boldsymbol{\gamma}_k||)^2, \text{ for } c_i \in \{1, ..., K\}$$

  where $\boldsymbol{\delta}_i$ is the feature vector for observation i.

- for each $k = 1, ..., K$, set

$$\boldsymbol{\gamma}_k := \frac{\sum_{i=1}^{N} \mathbf{1}_{\{c_i=k\}} \boldsymbol{\delta}_i}{\sum_{i=1}^{N} \mathbf{1}_{\{c_i=k\}}}.$$

We now present a simulation experiment that evaluates our proposed K-means partitioned geostatistical modelling framework on stationary and non-stationary simulated data.

3.4 Simulation

In this section, we evaluated performance of the proposed K-means partitioned geostatistical models. Performance was evaluated on two sets of simulated data; a set that exhibited stationarity in the covariance between observations, and a set that exhibited non-stationarity. The aim of the simulation experiment was to assess performance of these models in their

abilities to accurately predict, and account for spatial autocorrelation, in the values of a dependent variable. In particular, we used four measures of accuracy and a measure of residual spatial autocorrelation over a set of different scenarios. The measures of accuracy that we considered in each scenario were the root mean square error (RMSE, Equation 2.44) between observed and predicted values of the dependent variable, the mean absolute error (MAE, Equation 2.45) between observed and predicted values of the dependent variable, and the average estimation error of the covariance in logarithmic scale (COV, Equation 2.46). The measure of residual spatial autocorrelation that was calculated in each scenario was Moran's I, calculated from the residuals (Equation 2.47). Furthermore, we performed the experiments separately for the spatial case and the spatio-temporal case. For both cases, the models were determined by varying the number of partitions, K, as well as considering region specific and global coefficients and covariance parameters. Where possible, we used the same parameter values for each simulation out of preference. We fitted Matèrn covariance functions, with smoothness parameter $\nu = \frac{1}{2}$. We repeated each simulation 30 times to ensure that any patterns observed were not due to chance. The following sections outline the experimental designs and the simulation procedure.

3.4.1 Spatial simulation

A simulation experiment was conducted to evaluate the performance of K-means partitioned geostatistical models. The performance was evaluated on two sets of simulated data; one with a stationary spatial structure, and another with a non-stationary spatial structure. We repeated the simulation 30 times, for each set of data.

We randomly generated $N = 200$ longitude (s_{long}) and latitude (s_{lat}) values

from a unit square,

$$s_{\text{long}} \sim \text{U}(0,1),$$
$$s_{\text{lat}} \sim \text{U}(0,1).$$

Two datasets were considered for the simulation experiment. The first dataset, was one with a stationary spatial structure. We simulated a covariate, X_S, using the following equation to ensure that it was spatially correlated,

$$X_S = \frac{s_{\text{long}}}{2} + \frac{s_{\text{lat}}}{2} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 0.5).$$

A dependent variable with stationary spatial structure, $\boldsymbol{y}$, was simulated from,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \tag{3.7}$$

(Cameletti et al., 2011), where $\boldsymbol{X}\boldsymbol{\beta}$ is the linear combination of an intercept and the covariate, $\boldsymbol{\varepsilon}$ are the errors for the measurement process, and $\boldsymbol{\zeta}$ are the errors for the spatial process, that induced spatial autocorrelation in $\boldsymbol{y}$. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y} \sim \text{N}(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\boldsymbol{I} + \boldsymbol{\Sigma}), \tag{3.8}$$

where $\boldsymbol{\Sigma}$ was the spatial covariance matrix based on an Exponential covariance function,

$$\boldsymbol{\Sigma} = \sigma^2 \exp\left(\frac{-\boldsymbol{d}}{\psi}\right), \tag{3.9}$$

where $\boldsymbol{d}$ is a matrix with elements, d_{ij}, the Euclidean distance between location i and j. Here, the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \psi, \tau^2, \sigma^2)'$, were chosen to reflect a possible reality, for example, the temperature (in degrees Celsius) of a set of locations in a particular region at a particular time. We arbitrarily set $\boldsymbol{\beta} = (2, 1)'$, to represent the spatially varying mean in the context of the temperature example. This would produce a process where the average temperature across the region is two degrees Celsius and varies by a location dependent covariate X_S. We chose $\tau^2 = 0.1$ and $\sigma^2 = 1$ to ensure that the variability of the measurement process was less than that of the

spatial process in order to enhance the presence of spatial autocorrelation within $\boldsymbol{y}$. Finally, we set $\psi = 0.5$ to induce spatial autocorrelation. When choosing the parameter values, there was some degree of trial and error to ensure that the resulting data exhibited significant spatial autocorrelation, and stationarity. The stationary spatial structured data was sampled using Cholesky factorization (Algorithm 2, Rue & Held (2005)).

Figure 3.1 displays interpolated surface plots of X_S and $\boldsymbol{y}$ for one repetition of the simulation, and were produced to show the spatial autocorrelation within each of the variables. We see that $\boldsymbol{y}$ exhibits spatial autocorrelation, with clusters of higher values observed at the bottom left region as well as on the right. Clusters of lower values are observed at the top left of the plot. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}$ values. We calculated $I = 0.132$, with a p-value for the two-sided test for presence of spatial autocorrelation of $p < 2.2 \times 10^{-16}$, confirming the presence of significant spatial autocorrelation within the dependent variable. Similar observations were made when the simulation was repeated. We performed a test for the presence of non-stationarity using geographically weighted regression (Algorithm 1). We calculated p-values for the two-sided test for the presence of non-stationarity of $p = 0.015$ for the intercept, and $p = 0.99$ for the covariate, which suggests that there is no evidence for non-stationarity within the simulated data. This confirmed that the simulated dependent variable has a stationary spatial structure. Once again, similar observations were made when the simulation was repeated.

A dependent variable with non-stationary spatial structure, $\boldsymbol{y}$, was also simulated from Equations 3.7 and 3.9. In order to induce non-stationarity, we simulated a covariance matrix by partitioning the data into $K = 3$ subsets using the K-means algorithm (Algorithm 3) on the locations. We then simulated a covariate separately for each data subset. This was done, in addition to the partitioned covariance matrix, to induce non-stationarity within the dependent variable. The following equations were used to sim-

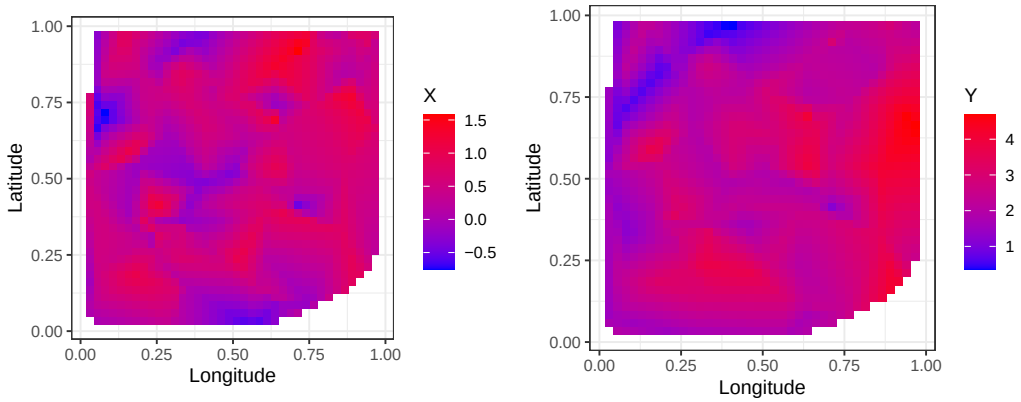


Figure 3.1: Interpolated surface plots of the covariate, top, and the simulated dependent variable, bottom, for one repeat of the simulation. Spatial autocorrelation is exhibited in the covariate as expected. Spatial autocorrelation is also evidenced for $\boldsymbol{y}$. There are clusters of high and low values throughout the surface plot for $\boldsymbol{y}$.

ulate the covariate, X_{NS},

$$X_{NS} \sim \mathrm{N}(b_k, g_k), \tag{3.10}$$

where $\boldsymbol{b} = (5, -2, 0)'$, and $\boldsymbol{g} = (0.5, 0.2, 0.7)'$, and $k = 1, 2, 3$ represents the data subset. We then calculated a covariance matrix for each data subset, according to,

$$(\boldsymbol{\Sigma})_{ij \in k} = \sigma_k^2 \exp\left(\frac{-d_{ij \in k}}{\psi_k}\right), \tag{3.11}$$

where $d_{ij \in k}$ is the Euclidean distance between location i and j, in partition $k = 1, 2, 3$. Like the stationary simulation, the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2)'$, were chosen to reflect the temperature (in degrees Celsius) of a set of locations in a particular region at a particular time. Once again, we set $\boldsymbol{\beta} = (2, 1)'$ to represent the spatially varying mean, which would produce a process where the average temperature across the region is two degrees Celsius and varies by a location dependent variable X_{NS}. We set $\boldsymbol{\tau}^2 = (0.1, 0.07, 0.04)'$ and $\boldsymbol{\sigma}^2 = (1, 0.7, 0.4)'$ to ensure that the variability of the measurement process was less than that of the spatial process and to enhance the presence of spatial autocorrelation within $\boldsymbol{y}$. Finally, $\boldsymbol{\psi} = (0.7, 0.4, 0.1)'$ to induce spatial autocorrelation. The parameter values for $\boldsymbol{\tau}^2$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\psi}$ were chosen arbitrarily and to ensure that the resulting data exhibited significant spatial autocorrelation, and non-stationarity. Different values where chosen for the parameters associated with each sub-region to induce non-stationarity. The non-stationary spatial structured data was sampled using Cholesky factorization (Algorithm 2, Rue & Held (2005)).

Figure 3.2 displays interpolated surface plots of X_{NS} and $\boldsymbol{y}$ for one repetition of the simulation, and were produced to show the spatial autocorrelation within each of the variables. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}$ values. We calculated $I = 0.279$, with a p-value for the two-sided test for presence of spatial autocorrelation of $p < 2.2 \times 10^{-16}$, confirming the presence of significant spatial autocorrelation within the dependent variable. Similar ob-

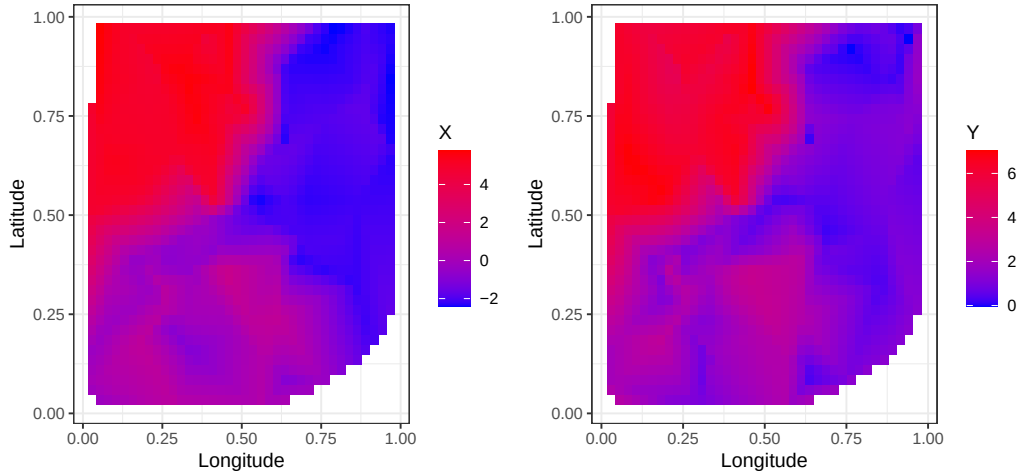


Figure 3.2: Interpolated surface plots of the covariate, top, and the simulated dependent variable, bottom, for one repeat of the simulation. Spatial autocorrelation is exhibited in the covariate as expected. Spatial autocorrelation is also evidenced for $\boldsymbol{y}$. There are clusters of high and low values throughout the surface plot for $\boldsymbol{y}$.

servations were made when the simulation was repeated. We performed a test for the presence of non-stationarity using geographically weighted regression (Algorithm 1). We calculated p-values for the two-sided test for the presence of non-stationarity of $p = 0$ for both the intercept and the covariate, which confirmed that the covariance of the dependent variable changes over space. Therefore, the simulated dependent variable was found to have a non-stationary spatial structure. Once again, similar observations were made when the simulation was repeated.

We wish to compare our proposed K-means partitioned geostatistical model over values of K ranging from $K = 1$ to $K = 5$, in order to account for spatial autocorrelation in, and make accurate predictions from, non-stationary spatially structured data. We note that when $K = 1$, the model becomes the standard, non-partitioned geostatistical model with stationary covariance function. In addition, we wish to determine the effect of including

region specific coefficients and covariance parameters on the measures of predictive accuracy and residual spatial autocorrelation. Therefore, we fitted nine models to each of the simulated sets of data. Model details are given in Table 3.1

Table 3.1: Details of the nine models fitted to both the stationary and non-stationary simulated data, for each repetition of the simulation.

Model	Equation	K	Parameters
1	$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}$	1	$\beta_0, \beta_1, \psi, \sigma^2, \tau^2$
2 – 5	$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k$	$2-5$	$\beta_0, \beta_1, \psi, \sigma^2, \tau^2$
6 – 9	$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k$	$2-5$	$\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\psi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2$

We first fitted Model 1, a standard, non-partitioned, geostatistical model with stationary Matèrn covariance function, to both sets of simulated data. Model 1 was defined by Equations 3.7 to 3.9, the same model used to simulate the stationary dataset. The parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \psi, \sigma^2, \tau^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_0, \beta_1 \sim \text{N}(0, 10), \;\; \psi \sim \text{IG}(3, 1), \;\; \sigma^2 \sim \text{IG}(3, 1), \;\; \tau^2 \sim \text{IG}(3, 1).$$

We then fitted Models 2 – 5 to both sets of simulated data. Models 2 – 5 were defined by our proposed K-means partitioned geostatistical model, Equations 3.2 and 3.4, where $K = 2, ..., 5$, respectively. Partitioning was performed using the K-means algorithm (Algorithm 3), based on the longitude and latitude of each locations. We assumed global model coefficients and covariance parameters for Models 2 – 5. For each model, the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \psi, \sigma^2, \tau^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_0, \beta_1 \sim \text{N}(0, 10), \;\; \psi \sim \text{IG}(3, 1), \;\; \sigma^2 \sim \text{IG}(3, 1) \;\; \tau^2 \sim \text{IG}(3, 1),$$

for $K = 2, ..., 5$.

Finally, we fitted Models 6 – 9 to both sets of simulated data. Models 6 – 9 were once again defined by our proposed K-means partitioned geostatistical model, Equations 3.5 to 3.6, where $K = 2, ..., 5$, respectively. Partitioning was again performed using the K-means algorithm (Algorithm 3), based on the longitude and latitude of each location. We assumed that the model coefficients and covariance parameters were sub-region specific for Models 6 – 9. For each model, the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_{01}, ...,\beta_{0K}, \beta_{11}, ..., \beta_{1K} \sim \text{N}(0, 10), \;\; \psi_1, ..., \psi_K \sim \text{IG}(3, 1),$$
$$\sigma_1^2, ..., \sigma_K^2 \sim \text{IG}(3, 1) \;\; \tau_1^2, ..., \tau_K^2 \sim \text{IG}(3, 1),$$

for $K = 2, ..., 5$.

We used MCMC to fit the models to both simulated sets of data and for each repetition. For each fitted model, two chains, each 100000 iterations, were generated of the parameter vector $\boldsymbol{\theta}$ for each dataset. We observed the chains converging to stationary distributions slowly and so 90000 (90%) of the iterations were discarded as warm-up. We thinned each chain by 5, to minimize autocorrelation in the posterior samples affording posterior draws of size 4000. Trace plots, density curves, and autocorrelation plots were checked to determine that the posterior samples converged to stationary distributions. For conciseness, we provided in Figures A.1 and A.2, diagnostic plots for Model 7 fitted to the first repetition of stationary and non-stationary simulated spatial data only. In addition to diagnostic plots, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, and these are displayed for each model fitted to the first repetition of stationary and non-stationary simulated spatial data in Tables A.1 – A.4. For each model fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary distributions, and appropriate exploration and mixing of the posterior distributions. Furthermore, $\hat{R}$ was found to be close to 1 for each parameter, indicating convergence (Brooks & Gelman, 1998).

We computed the posterior predictive distribution (Equation 2.43) for each model and obtained posterior distributions of fitted values. The posterior distributions of fitted values were used to calculate posterior distributions for RMSE, MAE, and Moran's I on the residuals, for each model fitted to each dataset. We also calculated the posterior average estimation error of the covariance in logarithmic scale, COV.

Table 3.2 displays the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9, averaged over each repetition of the simulated stationary spatial data and Figure 3.3 displays the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9, averaged over each repetition. Table 3.3 displays the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9 averaged over each repetition of the simulated non-stationary dataset and Figure 3.4 displays the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9, averaged over each repetition.

Table 3.2: Medians of the posterior distributions of RMSE, MAE, COV, and residual Moran's I for Models 1 – 9 averaged over each repetition of the simulated stationary spatial data.

Model	RMSE	MAE	COV	Moran's I
1	1.580	1.195	1.318	0.042
2	1.536	1.151	1.699	0.055
3	1.525	1.137	1.745	0.092
4	1.514	1.124	1.765	0.100
5	1.498	1.119	1.780	0.109
6	1.505	1.181	2.127	0.235
7	1.583	1.252	2.103	0.357
8	1.594	1.261	2.048	0.403
9	1.606	1.264	2.045	0.430

According to Table 3.2, we observed small differences in posterior median RMSE and MAE between each model, for each repetition. In general, Models 2 – 5 all had smaller posterior median RMSE and MAE compared to Model 1, the model used to generate the stationary data, and an existing methodology. This suggests that the models that partitioned the study region and assumed global parameters for the covariates and covariance matrix for each region were more accurate in prediction compared to an existing geostatistical methodology. Furthermore, and in general, the higher the number of partitions, the more accurate in prediction the models became. This trend was not observed for the models that assumed local effects of the covariates for each region. For Models 7 – 9, the predictive accuracy in terms of posterior median RMSE and MAE were markedly worse than that of Models 1 – 5. These results were further reflected in Figure 3.3. However, for Model 6, which partitioned the spatial domain into two regions and assumed local effects of the covariates for each region, the posterior median RMSE and MAE was lower than that of Model 1, and comparable to that of Models 2 – 5. This implies that allowing the covariate effects for both regions to differ in Model 6 improves predictive accuracy as well as not allowing them to differ. We also observed a somewhat decreasing trend in the amount of spread within the posterior distributions of RMSE and MAE for Models 1 – 5.

The posterior COV for Model 1 fitted to the stationary simulated data was the smallest of those for Models 1 – 9, for each repetition. This indicated that the parameters of the covariance function used to generate the data were estimated accurately. The posterior COV increased when the number of partitions increased, for both sets of models that treated the coefficients as global and local.

We observed an increasing trend in Moran's I, when the number of partitions increased. Furthermore, Moran's I was higher when the model assumed local parameters for each sub-region compared to global parameters. This trend was observed in all 30 simulations. It suggests that Models

1 – 5 were able to account for more spatial autocorrelation than Models 6 – 9.

We observed different results (Table 3.3) when the models were fitted to the non-stationary data. Figure 3.4 showed that the posterior median RMSE and MAE, averaged over each simulation, was lowest for Models 2 – 5, and 7. Model 7 represents the model that was used to generate the non-stationary data, where three partitions were used to split the domain and local parameters were estimated in each sub-region. Models 2 – 5 used two to five partitions, respectively, and assumed global coefficients for each sub-region. Each other model exhibited higher posterior median RMSE and MAE. This suggests that when spatial data exhibits non-stationarity, we can obtain better predictive accuracy (in terms of RMSE and MAE) when we partition the domain using the K-means algorithm, and estimate global parameters in each sub-region, or when we correctly specify the covariance matrix.

When compared to Model 1, which represents a traditional Matèrn covariance model, we observed lower posterior median RMSE and MAE for Models 2 – 5 ,and 7, averaged over each simulation. This suggests that fitting a model with a partitioned covariance structure that assumed global coefficients for each sub-region provides better predictive accuracy than a traditional methodology.

When the data had a non-stationary spatial structure, partitioning did not improve the accuracy of the estimation of the covariance matrix compared to the existing geostatistical methodology. The posterior COV for Model 1 fitted to the non-stationary simulated data was the smallest, suggesting that using a traditional geostatistical model that assumed global stationarity produced the most accurate covariance matrix.

We observed the same increasing trend in Moran's I, when the number of partitions increased, as we did with the stationary case. Moran's I was higher when the model assumed local parameters for each sub-region compared to global parameters. This trend was observed in all 30 sim-

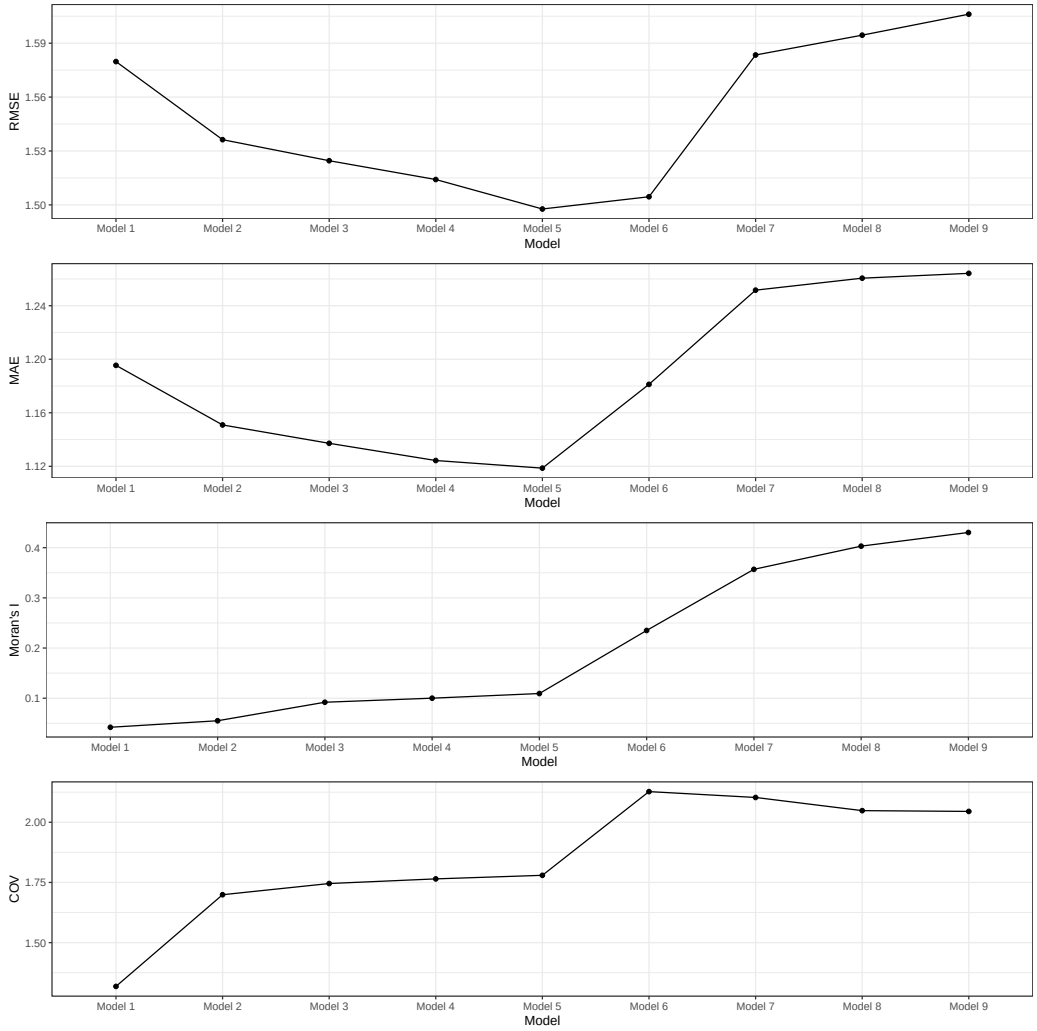


Figure 3.3: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for each model fitted to the stationary spatial data and averaged over each simulation repetition, where each model has different degrees of partitioning. Model 1 has no partitioning, Models 2 and 6 have two partitions, Models 3 and 7 have three partitions, Models 4 and 8 have four partitions, and Models 5 and 9 have five partitions. Further, Models 2 – 5 assume equal parameters across each sub-region, and Models 6 – 9 assume local parameters within each sub-region

Table 3.3: Medians of the posterior distributions of RMSE, MAE, COV, and residual Moran's I for Models 1 – 9 averaged over each repetition of the simulated non-stationary spatial data.

Model	RMSE	MAE	COV	Moran's I
1	1.309	1.011	1.184	0.031
2	1.281	0.988	1.359	0.073
3	1.271	0.987	1.392	0.106
4	1.257	0.960	1.423	0.105
5	1.254	0.961	1.429	0.119
6	1.298	1.022	1.491	0.248
7	1.279	0.992	1.507	0.381
8	1.343	1.066	1.554	0.443
9	1.361	1.072	1.601	0.477

ulations and it suggests that Models 1 – 5 were able to account for more spatial autocorrelation than Models 6 – 9. Model 1, the traditional Matèrn covariance model accounted for the spatial autocorrelation the best.

3.4.2 Spatio-temporal simulation

A simulation experiment was conducted to evaluate the performance of our proposed K-means partitioned Matèrn covariance function model on spatio-temporal data with a non-stationary spatial structure. We also compared our models to that of a traditional Matèrn covariance function model. The performance was evaluated on two sets of simulated data; one with a stationary spatial structure, and another with a non-stationary spatial structure. For both datasets, we induced spatial and temporal autocorrelation in a dependent variable, and assumed that there was no interaction between space and time. We repeated the simulation 30 times, for each set of data.

As with the spatial simulation, we randomly generated $N = 200$ longitude

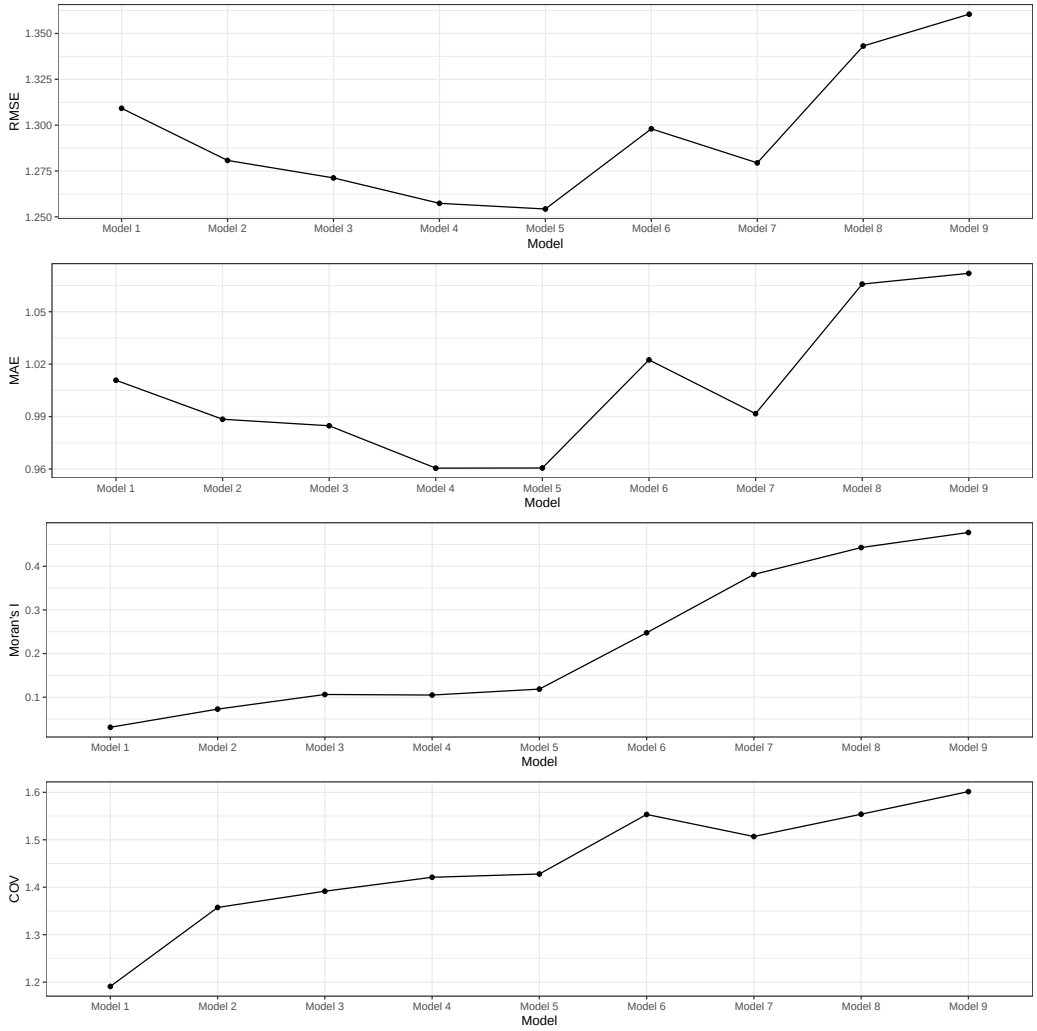


Figure 3.4: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for each model fitted to the non-stationary spatial data and averaged over each simulation repetition, where each model has different degrees of partitioning. Model 1 has no partitioning, Models 2 and 6 have two partitions, Models 3 and 7 have three partitions, Models 4 and 8 have four partitions, and Models 5 and 9 have five partitions. Further, Models 2 – 5 assume equal parameters across each sub-region, and Models 6 – 9 assume local parameters within each sub-region

(s_{long}) and latitude (s_{lat}) values from a unit square,

$$s_{\text{long}} \sim \text{U}(0,1),$$
$$s_{\text{lat}} \sim \text{U}(0,1).$$

We simulated a covariate, X_S, for $T = 5$ time points, using the following equation to ensure that it was spatially correlated,

$$X_S = \frac{s_{\text{long}}}{2} + \frac{s_{\text{lat}}}{2} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 0.5).$$

A dependent variable with stationary spatio-temporal structure, $\boldsymbol{y}_t$ for each time point $t = 1, ..., 5$, was simulated from,

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\zeta}_t + \boldsymbol{\varepsilon}_t, \tag{3.12}$$

(Cameletti et al., 2011), where $\boldsymbol{X}_t\boldsymbol{\beta}$ is the linear combination of an intercept and the covariate at time t, $\boldsymbol{\varepsilon}_t$ are the errors for the measurement process, at time t and $\boldsymbol{\zeta}_t$ are the errors for the spatio-temporal process at time t, that induce spatial and temporal autocorrelation in $\boldsymbol{y}_t$. We modelled the spatio-temporal process as,

$$\boldsymbol{\zeta}_t = \rho\boldsymbol{\zeta}_{\boldsymbol{t-1}} + \boldsymbol{\omega}_t, \tag{3.13}$$

where ρ is a temporal correlation coefficient. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y}_t \sim \text{N}(\boldsymbol{X}_t\boldsymbol{\beta}, \tau^2\boldsymbol{I} + \boldsymbol{\Sigma}_t), \tag{3.14}$$

where $\boldsymbol{\Sigma}_t$ is the spatio-temporal covariance matrix based on Exponential covariance function,

$$\boldsymbol{\Sigma}_t = \frac{\sigma^2}{1-\rho^2} \exp\left(\frac{-\boldsymbol{d}}{\psi}\right), \tag{3.15}$$

where $\boldsymbol{d}$ is a matrix with elements, d_{ij}, the Euclidean distance between location i and j. Staying in line with the spatial simulation experiments, the parameters for this spatio-temporal simulation $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho, \psi, \tau^2, \sigma^2)'$, were

chosen to reflect a reality, such as the temperature (in degrees Celsius) of a set locations in a particular region at a particular time. We set $\boldsymbol{\beta} = (2, 1)'$ to represent the spatially varying mean in the temperature example. Doing this would produce a process where the average temperature across the region is two degrees Celsius and varies by a location dependent variable X_S. We selected $\tau^2 = 0.1$ and $\sigma^2 = 1$ to ensure that the measurement process had less variability than the spatial process and to enhance the presence of spatial autocorrelation within $\boldsymbol{y}_t$. The temporal autocorrelation parameter was set to $\rho = 0.7$ to induce moderate temporal autocorrelation, and we chose $\psi = 0.5$ to induce spatial autocorrelation. When choosing specific covariance parameter values, there was some degree of trial and error to ensure that the resulting data exhibited significant spatial autocorrelation, and stationarity. The stationary spatial structured data was sampled using Cholesky factorisation (Algorithm 2, Rue & Held (2005)).

Figure 3.5 displays interpolated surface plots of X_S and $\boldsymbol{y}_t$ for $t = 1$ and for one repetition of the simulation, and were produced to show the spatial autocorrelation within each of the variables. For each time point t, we see that $\boldsymbol{y}_t$ exhibits spatial autocorrelation, with clusters of higher values observed at the upper right region and middle left region, while clusters of lower values were observed at the bottom right region. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}_t$ values for each t. We calculated $I_{t=1} = 0.125$, $I_{t=2} = 0.123$, $I_{t=3} = 0.147$, $I_{t=4} = 0.138$, and $I_{t=5} = 0.116$, with p-values for the two-sided test for presence of spatial autocorrelation of $p < 1 \times 10^{-15}$ for each t, which confirmed the presence of significant spatial autocorrelation within the dependent variable. Similar observations were made when the simulation was repeated. We also performed a test for the presence of non-stationarity using geographically weighted regression (Algorithm 1). All calculated p-values for the two-sided test for the presence of non-stationarity using the covariate were greater than 0.1, which suggested that there was no evidence for non-stationarity within the simulated data. This confirmed that

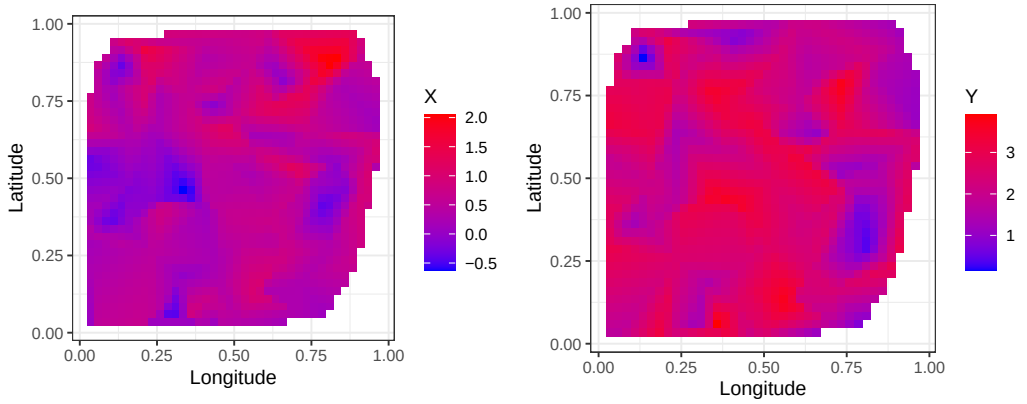


Figure 3.5: Interpolated surface plots of the covariate, left, and the simulated dependent variable, right, for $t = 1$. Spatial autocorrelation is exhibited in the covariate as expected. Spatial autocorrelation is also evidenced for $\boldsymbol{y}_1$. There are clusters of high and low values throughout the surface plot for $\boldsymbol{y}_1$.

the simulated dependent variable had a stationary spatial structure. Once again, similar observations were made when the simulation was repeated. A dependent spatio-temporal variable with non-stationary spatial structure, $\boldsymbol{y}_t$, was also simulated from Equations 3.12 and 3.15. In order to induce non-stationarity within the dependent variable, we simulated a spatial covariance matrix by partitioning the data into $K = 3$ subsets using the K-means algorithm (Algorithm 3) on the locations, independent of time. We then simulated a covariate separately for each data subset. This was done, in addition to the partitioned covariance matrix, to induce non-stationarity within the dependent variable. The following equations were used to simulate the covariate, X_{NS},

$$X_{NS} \sim \mathrm{N}(b_k, g_k), \tag{3.16}$$

where $\boldsymbol{b} = (5, -2, 0)'$, and $\boldsymbol{g} = (0.5, 0.2, 0.7)'$, and $k = 1, 2, 3$ represent the

data subset. We then calculated a spatial covariance matrix for each data subset, independent of time,

$$(\boldsymbol{\Sigma}_t)_{ij\in k} = \frac{\sigma_k^2}{1-\rho^2} \exp\left(\frac{-d_{ij\in k}}{\psi_k}\right), \tag{3.17}$$

where $d_{ij\in k}$ is the Euclidean distance between location i and j, in partition $k = 1, 2, 3$. The parameters for this simulation, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \rho, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2)'$, were chosen to reflect the same reality as the stationary spatio-temporal simulation. Once again, we set $\boldsymbol{\beta} = (2, 1)'$ to represent the spatially varying mean, which would produce a process where, for example, the average temperature across the region is two degrees Celsius and varies by a location dependent variable X_{NS}. We chose $\boldsymbol{\tau}^2 = (0.1, 0.07, 0.04)'$ and $\sigma^2 = (1, 0.7, 0.4)'$ to ensure that the measurement process had less variability than the spatial process and to enhance the presence of spatial autocorrelation within $\boldsymbol{y}_t$. The temporal autocorrelation parameter was set to $\rho = 0.7$ to induce moderate temporal autocorrelation, and we chose $\boldsymbol{\psi} = (0.7, 0.4, 1)'$ to induce spatial autocorrelation. When choosing $\boldsymbol{\tau}^2, \boldsymbol{\sigma}^2$, and $\boldsymbol{\psi}$, there was some degree of trial and error and to ensure that the resulting data exhibited significant spatial autocorrelation and non-stationarity. The non-stationary spatial structured data was sampled using Cholesky factorisation (Algorithm, 2 Rue & Held (2005)).

Figure 3.6 displays interpolated surface plots of X_{NS} and $\boldsymbol{y}_t$ for $t = 1$ for one repetition of the simulation, and were produced to show the spatial autocorrelation within each of the variables. For each time point t, we see that the dependent variable exhibits spatial autocorrelation, with clusters of higher values observed on the left and right regions and clusters of lower values observed in the middle region. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}_t$ values for each t. We calculated $I_{t=1} = 0.0288$, $I_{t=2} = 0.139$, $I_{t=3} = 0.197$, $I_{t=4} = 0.174$, and $I_{t=5} = 0.0958$, with p-values for the two-sided test for presence of spatial autocorrelation of $p < 0.05$ for each t, which confirmed the presence of significant spatial autocorrelation within the dependent

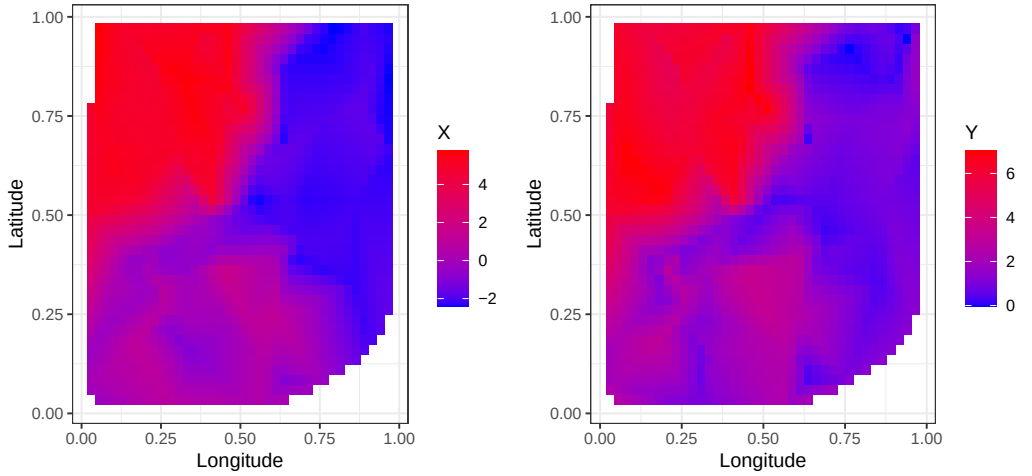


Figure 3.6: Interpolated surface plots of the covariate, top, and the simulated dependent variable, bottom. Spatial autocorrelation is exhibited in the covariate as expected. Spatial autocorrelation is also evidenced for $\boldsymbol{y}$. There is a general upward diagonal trend, with higher values observed at the top right of the plot, and lower values observed at the bottom left of the plot.

variable Similar observations were made when the simulation was repeated. We performed a test for the presence of non-stationarity using geographically weighted regression (Algorithm 1), which found that all calculated p-values for the two-sided test for the presence of non-stationarity using the covariate were approximately equal to 0, which suggested that there was evidence for significant non-stationarity within the simulated data. This confirmed that the simulated dependent variable had a non-stationary spatial structure. The same results were found when the simulation was repeated.

Like the spatial simulation, we wish to compare our proposed K-means partitioned geostatistical model over values of K ranging from $K = 1$ to $K = 5$, in order to account for spatial autocorrelation in, and make accurate predictions from, non-stationary spatio-temporally structured data.

Again, we note that when $K = 1$, the model becomes the standard, non-partitioned geostatistical model with stationary covariance function. In addition, we wish to determine the effect of including local parameters in the model on the measures of predictive accuracy and residual spatial autocorrelation. As with the spatial simulation, we fitted nine models to each of the simulated sets of spatio-temporal data. The models are detailed in Table 3.4.

Table 3.4: Details of the nine models fitted to both the stationary and non-stationary simulated data.

Model	Equation	K	Parameters
1	$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t + \boldsymbol{\varepsilon}_t$	1	$\beta_0, \beta_1, \psi, \rho, \sigma^2, \tau^2$
2 – 5	$\boldsymbol{y}_{kt} = \boldsymbol{X}_{kt}\boldsymbol{\beta} + \boldsymbol{Z}_{kt} + \boldsymbol{\varepsilon}_{kt}$	$2-5$	$\beta_0, \beta_1, \psi, \rho, \sigma^2, \tau^2$
6 – 9	$\boldsymbol{y}_{kt} = \boldsymbol{X}_{kt}\boldsymbol{\beta}_k + \boldsymbol{Z}_{kt} + \boldsymbol{\varepsilon}_{kt}$	$2-5$	$\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\psi}, \rho, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2$

We first fitted Model 1, the spatio-temporal equivalent to the first model fitted in the spatial simulation. This model is a non-partitioned geostatistical model with stationary Matèrn covariance function model, and was fitted to both sets of simulated data. Model 1 is defined by Equations 3.12 to 3.15, the same model used to simulate the dataset. The parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \psi, \rho, \tau^2, \sigma^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_0, \beta_1 \sim \text{N}(0, 10),\ \psi \sim \text{IG}(3, 1),\ \rho \sim \text{U}(-1, 1),\ \sigma^2 \sim \text{IG}(3, 1),\ \tau^2 \sim \text{IG}(3, 1).$$

Next, we fitted Models 2 – 5 to both sets of simulated data. Models 2 – 5 were defined by our proposed K-means partitioned geostatistical model for spatio-temporal data, given by Equations 3.5 and 3.6, where $K = 2, ..., 5$, respectively. Partitioning was performed using the K-means algorithm, given by Algorithm 3, based on the longitude and latitude for each location. We assumed global model coefficients and covariance parameters for Models 2 – 5, inline with the spatial simulation. For each model,

the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \psi, \rho, \tau^2, \sigma^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_0, \beta_1 \sim \mathrm{N}(0, 10),\ \psi \sim \mathrm{IG}(3, 1),\ \rho \sim \mathrm{U}(-1, 1),\ \sigma^2 \sim \mathrm{IG}(3, 1),\ \tau^2 \sim \mathrm{IG}(3, 1).$$

Finally, we fitted Models 6 – 9 to both sets of simulated data. Models 6 – 9 were defined in the same way as Models 2 – 5, with the exception that all model parameters except the temporal correlation coefficient, ρ, were assumed to be sub-region specific,. For each model, the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \rho, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_{01}, ..., \beta_{0K}, \beta_{11}, ..., \beta_{1K} \sim \mathrm{N}(0, 10),\ \ \psi_1, ..., \psi_K \sim \mathrm{IG}(3, 1),$$
$$\rho \sim \mathrm{U}(-1, 1),\ \ \sigma_1^2, ..., \sigma_K^2 \sim \mathrm{IG}(3, 1),\ \ \tau_1^2, ..., \tau_K^2 \sim \mathrm{IG}(3, 1).$$

We used MCMC to fit the models to both simulated sets of data and for each repetition. For each fitted model, two chains, each 100000 iterations, were generated of the parameter vector $\boldsymbol{\theta}$ for each dataset. We observed the chains converging to stationary distributions slowly and so 90000 (90%) of the iterations were discarded as warm-up. We thinned each chain by 5, to minimize autocorrelation in the posterior samples affording posterior draws of size 4000. Trace plots, density curves, and autocorrelation plots were checked to determine that the posterior samples converged to stationary distributions. For conciseness, we provided in Figures A.3 and A.4, diagnostic plots for Model 7 fitted to the first repetition of stationary and non-stationary simulated spatial data only. In addition to diagnostic plots, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, and these are displayed for each model fitted to the first repetition of stationary and non-stationary simulated spatial data in Tables A.5 – A.8. For most models fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary distributions, and appropriate exploration and mixing of the posterior distributions. Furthermore, $\hat{R}$ was found to be close to 1 for most parameters, indicating convergence (Brooks & Gelman, 1998).

We computed the posterior predictive distribution (Equation 2.43) for each model and obtained posterior distributions of fitted values. The posterior distributions of fitted values were used to calculate posterior distributions for RMSE, MAE, and Moran's I on the residuals, for each model fitted to each dataset. In addition, we calculated the posterior average estimation error of the covariance in logarithmic scale, COV. Figure 3.7 and Table 3.5 display the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9 for the stationary spatio-temporal dataset. Figure 3.8 and Table 3.6 display the medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for Models 1 – 9 for the non-stationary spatio-temporal dataset.

From Table 3.5 and Figure 3.7, we observed that, on average, Model's 4 and 5 had the lowest median posterior RMSE and MAE when fitted to the stationary spatio-temporal data. This indicated that these models provided the best predictive accuracy of all nine models fitted to the data. This is despite the fact that Model 1 had a larger median posterior RMSE and MAE and represented a standard Matèrn covariance model and was the data generation model. Models 4 and 5 were models that partitioned the domain into four and five sub-regions, respectively, and assumed global coefficients in each.

In general, the four models that assumed local coefficients, Models 6 – 9, had higher posterior median RMSE and MAE averaged over each repetition. This indicates that when the simulated data is stationary, a model that assumes global coefficients has better predictive accuracy than a model that assumes local coefficients for each partitioned sub-region.

The posterior COV for Model 1 fitted to the stationary data was the highest out of all nine models, averaged over each partition. This was unexpected, since Model 1 fitted a correctly specified covariance matrix to the data. All partition models fitted to the stationary data had relatively equal average posterior COV, indicating that the covariance matrix was more accurately estimated when we partitioned the spatial domain, regardless of assuming

Table 3.5: Medians of the posterior distributions of RMSE, MAE, COV, and residual Moran's I for Models 1 – 9 fitted to the first repetition of the simulated stationary spatio-temporal data.

Model	RMSE	MAE	COV	Moran's I				
				$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
1	2.360	1.974	2.707	0.199	0.227	0.244	0.235	0.297
2	2.447	2.015	2.041	0.180	0.189	0.201	0.185	0.213
3	2.450	2.016	1.870	0.186	0.200	0.208	0.204	0.226
4	2.270	1.851	1.703	0.169	0.184	0.193	0.186	0.213
5	2.220	1.804	1.725	0.193	0.200	0.207	0.201	0.224
6	2.551	2.120	2.011	0.194	0.195	0.208	0.196	0.232
7	2.477	2.050	1.822	0.175	0.185	0.194	0.196	0.208
8	2.596	2.153	1.965	0.209	0.211	0.224	0.224	0.248
9	2.711	2.247	2.012	0.204	0.209	0.215	0.212	0.231

global or local parameters in each sub-region.

In general, we observed larger median Moran's I values for the partitioned models that assumed local parameters than the partitioned models that assumed global parameters, for each t and averaged over each repetition. This suggests that when partitioned models that assume global parameters for each sub-region are fitted to stationary data, they are better at accounting for spatial autocorrelation than when local parameters were assumed. For most values of t, the median Moran's I for Model 1 (the traditional Matèrn model) fitted to the stationary data was the largest, averaged over each repetition. This suggests that fitting the models with a partitioned spatio-temporal covariance structure accounts for spatial autocorrelation the better than this existing methodology.

Unlike the spatial simulation, we observed similar results when we fitted the models to the non-stationary data. From Table 3.6 and Figure 3.8, we observed that, on average, Model's 1 and 5 had the lowest me-

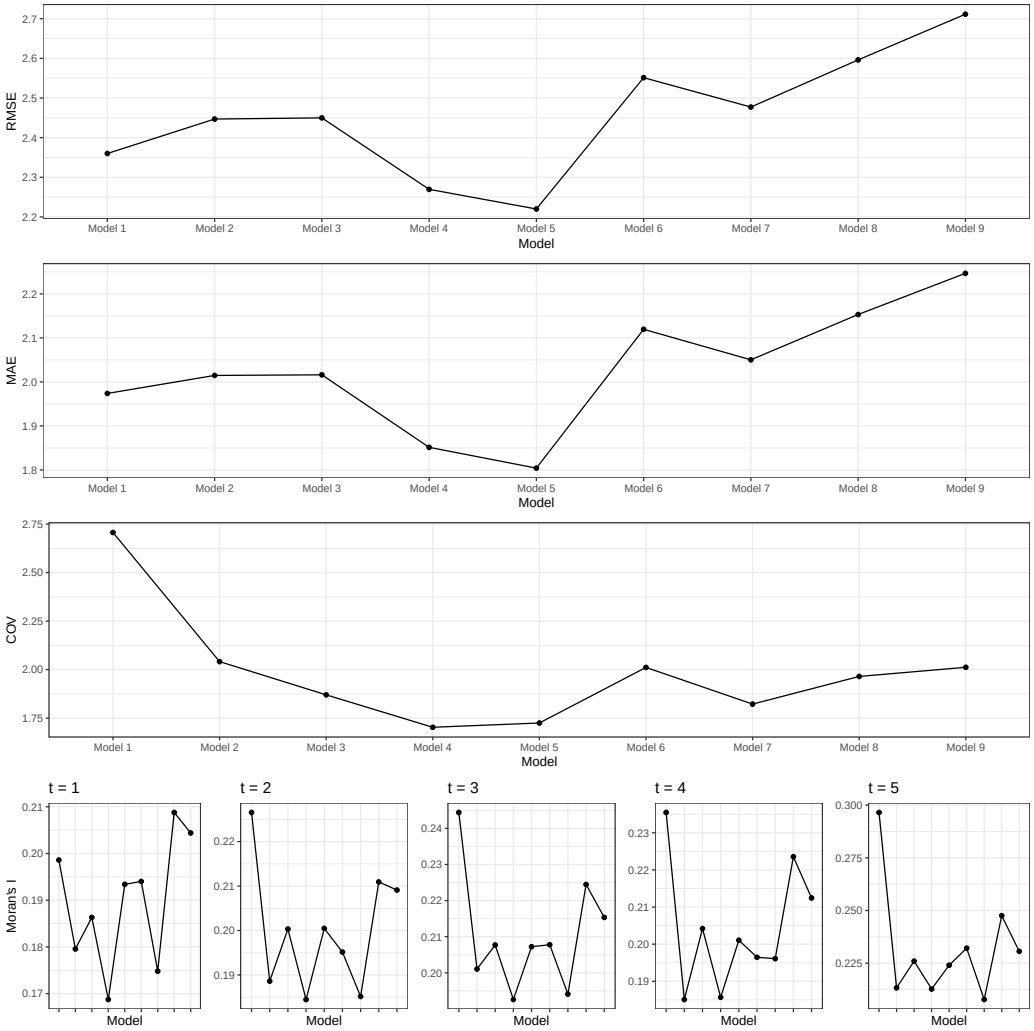


Figure 3.7: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for each model fitted to the stationary spatio-temporal data, where each model has different degrees of partitioning. Model 1 has no partitioning, Models 2 and 6 have two partitions, Models 3 and 7 have three partitions, Models 4 and 8 have four partitions, and Models 5 and 9 have five partitions. Further, Models 2 – 5 assume equal parameters across each sub-region, and Models 6 – 9 assume local parameters within each sub-region

dian posterior RMSE and MAE when fitted to the non-stationary spatio-temporal data. This indicated that these models provided the best predictive accuracy of all nine models fitted to the data. This was unusual, since Model 7 was used to generate the non-stationary data. In this case, Model 7 afforded relatively large median posterior RMSE and MAE, indicating worse predictive accuracy.

Table 3.6: Medians of the posterior distributions of RMSE, MAE, COV, and residual Moran's I for Models 1 – 9 fitted to the first repetition of the simulated non-stationary spatio-temporal data.

Model	RMSE	MAE	COV	Moran's I				
				$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
1	1.954	1.601	3.423	0.172	0.176	0.098	0.086	0.171
2	2.145	1.721	2.331	0.143	0.139	0.117	0.109	0.142
3	2.205	1.770	2.442	0.144	0.147	0.131	0.125	0.150
4	2.067	1.658	2.089	0.143	0.148	0.117	0.113	0.144
5	1.877	1.502	1.713	0.107	0.121	0.084	0.078	0.113
6	2.158	1.736	2.652	0.164	0.152	0.142	0.133	0.165
7	2.039	1.652	1.772	0.182	0.177	0.146	0.141	0.179
8	2.039	1.621	2.223	0.159	0.175	0.147	0.133	0.154
9	1.969	1.573	1.957	0.154	0.164	0.136	0.125	0.153

The posterior COV for Model 5 and was the lowest, averaged over each repetition of the simulation. This indicated that when the number of partitions used was five, and the parameters were assumed global within each sub-region, the covariance matrix was estimated accurately. The posterior COV for Model 7 was also lower than that of the other models, which is not surprising since Model 7 represents the data generating model. The posterior COV for Model 1, which represents the standard Matèrn model fitted to the non-stationary data was the highest out of all nine models, averaged over each partition. This was expected in this simulation ex-

periment, since Model 1 does not correctly account for non-stationarity in the covariance matrix. Further, it suggests that partitioning the covariance matrix can produce more accurate estimates of the covariance matrix when data has non-stationary spatial structure.
We observed unusual patterns in median Moran's I values. In general, median posterior Moran's I on the residuals were higher for the partitioned models that assumed local parameters than the partitioned models that assumed global parameters, for each t and averaged over each repetition. This suggests that when partitioned models that assume global parameters for each sub-region are fitted to stationary data, they are better at accounting for spatial autocorrelation than when local parameters were assumed. For most values of t, the median Moran's I for Models 1, 2, and 5 fitted to the stationary data was the lowest, averaged over each repetition.

3.5 Case study

We performed an exploratory case study, in which we fitted several K-means partitioned geostatistical models to the 2013 New Zealand particulate matter (PM10) concentration data described in Section 2.7.1. The aim was to find the best model to predict PM10 concentration across New Zealand, in terms of predictive accuracy measures RMSE, and MAE, as well as in terms of the accuracy of estimation of the covariance matrix, while also accounting for spatial autocorrelation and non-stationarity. When a best model was chosen, we used it to produce an interpolated predictive map for particulate matter concentration, using covariate observations where monitoring stations were not placed. Limited covariates were available to estimate the models, with only temperature (in °C) and wind speed (in m/s) considered. Temporal variation was not considered for this case study.
Mean PM10 concentration was observed at 40 locations across New Zealand

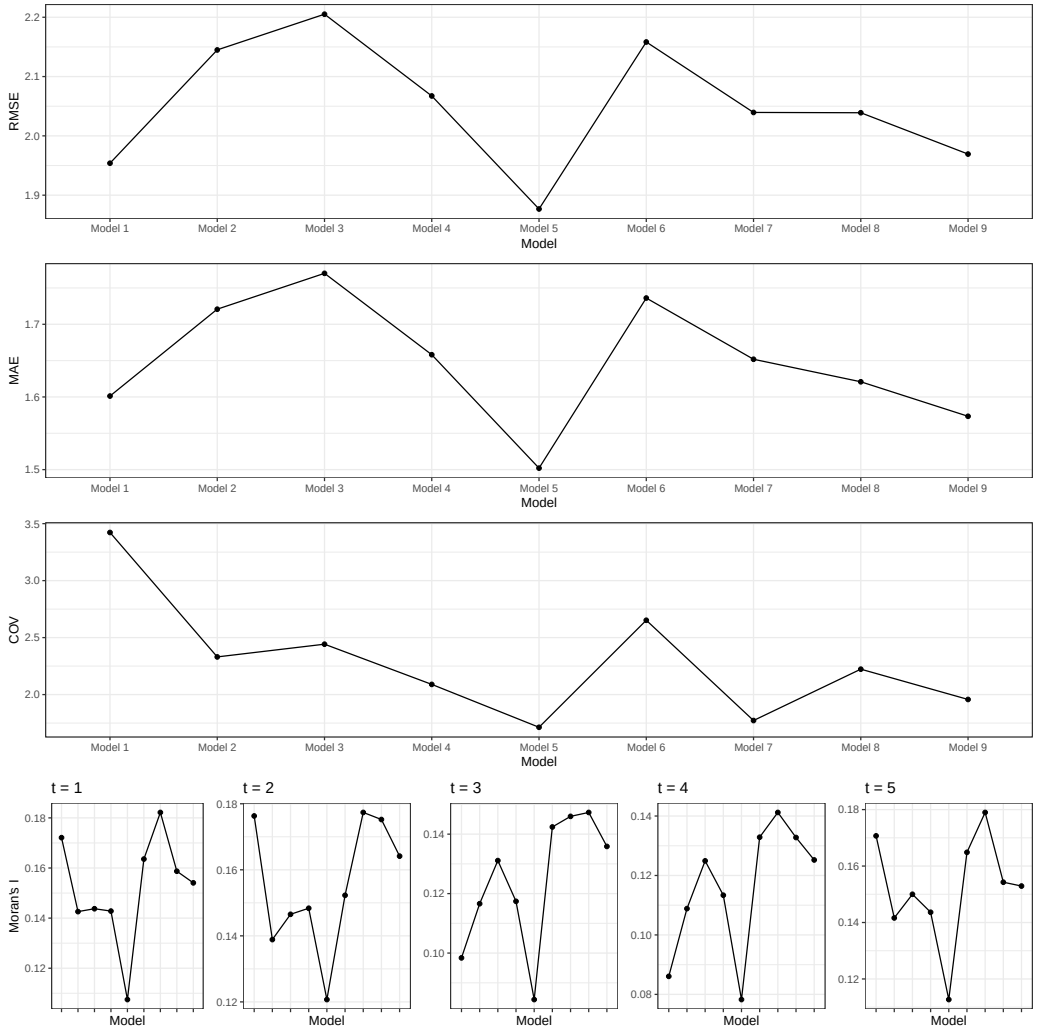


Figure 3.8: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I for each model fitted to the non-stationary spatio-temporal data, where each model has different degrees of partitioning. Model 1 has no partitioning, Models 2 and 6 have two partitions, Models 3 and 7 have three partitions, Models 4 and 8 have four partitions, and Models 5 and 9 have five partitions. Further, Models 2 – 5 assume equal parameters across each sub-region, and Models 6 – 9 assume local parameters within each sub-region

for the year 2013, denoted by $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))'$, shown in Figure 2.2. Significant spatial autocorrelation was identified across the study region, as evidenced by this figure. This was confirmed by Moran's I, which was calculated as $I = 0.3577$ with a corresponding p-value for the two-sided test for the presence of spatial autocorrelation of 3.23×10^{-8}. Furthermore, significant non-statitonarity was identified. A p-value for the two-sided test for non-stationarity was calculated to be 0.03. This provided motivation for a partitioned geostatistical model.

The partitioned geostatistical model described by Equations 3.2 and 3.4 was fitted to the natural log of the New Zealand PM10 concentrations. We partitioned the spatial domain into K distinct sub-regions using the K-means algorithm defined in Algorithm 3. We determined the models by varying K, ranging from $K = 1$ to $K = 3$, to ensure that there was a reasonable number of observations within each partition. In addition, we fitted the model assuming global coefficients as well as region specific coefficients. We summarised the models fitted to the data in Table 3.7.

Table 3.7: Details of the five models fitted to the New Zealand particulate matter data.

Model	Equation	K	Parameters
1	$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}$	1	$\beta_0, \beta_1, \beta_2, \psi, \sigma^2, \tau^2$
2, 3	$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k$	2, 3	$\beta_0, \beta_1, \beta_3\psi, \sigma^2, \tau^2$
4, 5	$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k$	2, 3	$\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\psi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}^2$

By assuming conditional independence between the regions, the model equation for $\boldsymbol{y}$ is given by,

$$\log \boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\zeta}_k + \boldsymbol{\varepsilon}_k, \tag{3.18}$$

where $\boldsymbol{X}_k = (\boldsymbol{1}_{n_k}, \boldsymbol{x}_{1k}, \boldsymbol{x}_{2k})$ is the $n_k \times 3$ design matrix of covariates, with $\boldsymbol{x}_{1k}$ being the temperature observed at locations in partition k, and $\boldsymbol{x}_{2k}$ being the wind speed observed at locations in partition k. Here, $\boldsymbol{\beta}_k$ are

the corresponding coefficients, with $\boldsymbol{\beta}_k = \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ for Models 1 – 3, and $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \beta_{2k})'$ for Models 4 and 5, for $k = 1, ..., K$, $K \in \{1, 2, 3\}$. Furthermore, $\boldsymbol{\zeta}_k$ is the spatial error term, assumed to follow $\boldsymbol{\zeta}_k \sim \mathrm{N}(\mathbf{0}, \sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k))$, with region specific covariance matrix, depending on covariance function parameters $\boldsymbol{\phi}_k$. Finally, $\boldsymbol{\varepsilon}_k$ is the measurement error term, assumed to follow $\boldsymbol{\varepsilon}_k \sim \mathrm{N}(\mathbf{0}, \tau_k^2 \boldsymbol{I}_{n_k})$, with region specific nugget variance τ_k^2.

We chose to model the covariance matrix of using the Matèrn covariance function with smoothness parameter $\nu = 0.5$, which has an exponential closed form,

$$\sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k) = \sigma_k^2 \exp\left(- \frac{\boldsymbol{d}_k}{\psi_k} \right), \tag{3.19}$$

where $\boldsymbol{d}_k$ is the matrix of Euclidean distances between pairs of observations in region k, and ψ_k is the strength of spatial correlation. We now describe each model in more detail.

3.5.1 Models

Model 1

Model 1 is given by Equation 3.18 when $K = 1$. In other words, it is the standard non-partitioned geostatistical model. The model assumed that the PM10 data was stationary and estimated a single covariance function for the entire study region. The data likelihood is given by,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}) &\propto |\sigma^2 \boldsymbol{R}(\psi) + \tau^2 \boldsymbol{I}|^{-\frac{1}{2}} \\ &\times \exp\left\{ -\frac{1}{2} (\log \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\sigma^2 \boldsymbol{R}(\psi) + \tau^2 \boldsymbol{I})^{-1} (\log \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}, \end{aligned} \tag{3.20}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \psi)'$ and $R(\psi)$ is defined by Equation 3.19 with $\psi_k = \psi$. The parameters $\boldsymbol{\theta}$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta_0, \beta_1, \beta_2 \sim \mathrm{N}(0, 10), \;\; \psi \sim \mathrm{IG}(2, 1), \;\; \sigma^2 \sim \mathrm{IG}(2, 1), \;\; \tau^2 \sim \mathrm{IG}(2, 1).$$

This afforded the posterior distribution,

$$
\begin{aligned}
f(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}) &\propto |\sigma^2\boldsymbol{R}(\psi)+\tau^2\boldsymbol{I}|^{-\frac{1}{2}} \\
&\times \exp\left\{-\frac{1}{2}(\log\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma^2\boldsymbol{R}(\psi)+\tau^2\boldsymbol{I})^{-1}(\log\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}'(10\boldsymbol{I}_3)^{-1}\boldsymbol{\beta})\right)\times(\sigma^2)^{-3}\exp\left(-\frac{1}{\sigma^2}\right) \\
&\times (\tau^2)^{-3}\exp\left(-\frac{1}{\tau^2}\right)\times(\psi^2)^{-3}\exp\left(-\frac{1}{\psi^2}\right).
\end{aligned}
\tag{3.21}
$$

Models 2 and 3

Models 2 and 3 are given by Equation 3.18 when $K=2$ and $K=3$, respectively, and assumed that the covariate coefficients were equal within each partition while the covariance parameters were not. The data likelihoods for both models are given by,

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{X}) &\propto |\boldsymbol{\Sigma}+\mathbf{T}|^{-\frac{1}{2}} \\
&\times \exp\left\{-\frac{1}{2}(\log\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{\Sigma}+\mathbf{T})^{-1}(\log\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\},
\end{aligned}
\tag{3.22}
$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\tau}^2,\boldsymbol{\psi})'$, and $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are block-diagonal matrices with main diagonal matrices $\sigma_k^2\boldsymbol{R}(\psi_k)$ and $\tau_k^2\boldsymbol{I}_{n_k}$, respectively, and $k=1,...,K$, with $K=2$ for Model 2, and $K=3$ for Model 3.The parameters $\boldsymbol{\theta}$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$
\boldsymbol{\beta}\overset{iid}{\sim}\mathrm{N}(0,10),\ \ \boldsymbol{\psi}\overset{iid}{\sim}\mathrm{IG}(2,1),\ \ \boldsymbol{\sigma}^2\overset{iid}{\sim}\mathrm{IG}(2,1),\ \ \boldsymbol{\tau}^2\overset{iid}{\sim}\mathrm{IG}(2,1).
$$

This afforded the posterior distribution given by,

$$
\begin{aligned}
f(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}) &\propto |\boldsymbol{\Sigma}+\mathbf{T}|^{-\frac{1}{2}} \\
&\times \exp\left\{ -\frac{1}{2}(\log \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{\Sigma}+\mathbf{T})^{-1}(\log \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}'(10\boldsymbol{I}_3)^{-1}\boldsymbol{\beta})\right) \times \prod_{k=1}^{K}(\sigma_k^2)^{-3}\exp\left(-\frac{1}{\sigma_k^2}\right) \\
&\times \prod_{k=1}^{K}(\tau_k^2)^{-3}\exp\left(-\frac{1}{\tau_k^2}\right) \times \prod_{k=1}^{K}(\psi_k^2)^{-3}\exp\left(-\frac{1}{\psi_k^2}\right).
\end{aligned}
\tag{3.23}
$$

Models 4 and 5

Models 4 and 5 are given by Equation 3.18 when $K=2$ and $K=3$, respectively, and assume that the model parameters are not equal within each partition. The data likelihoods for both models are given by,

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{X}) &\propto |\boldsymbol{\Sigma}+\mathbf{T}|^{-\frac{1}{2}} \\
&\times \exp\left\{ -\frac{1}{2}(\log \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{\Sigma}+\mathbf{T})^{-1}(\log \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\},
\end{aligned}
\tag{3.24}
$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\tau}^2,\boldsymbol{\psi})'$, and $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are block-diagonal matrices with main diagonal matrices $\sigma_k^2\boldsymbol{R}(\psi_k)$ and $\tau_k^2\boldsymbol{I}_{n_k}$, respectively, and $k=1,...,K$, with $K=2$ for Model 2, and $K=3$ for Model 3.The parameters $\boldsymbol{\theta}$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$
\boldsymbol{\beta} \overset{iid}{\sim} \mathrm{N}(0,10),\ \ \boldsymbol{\psi} \overset{iid}{\sim} \mathrm{IG}(2,1),\ \ \boldsymbol{\sigma}^2 \overset{iid}{\sim} \mathrm{IG}(2,1),\ \ \boldsymbol{\tau}^2 \overset{iid}{\sim} \mathrm{IG}(2,1).
$$

This afforded the posterior distribution given by,

$$
\begin{aligned}
f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) &\propto |\boldsymbol{\Sigma} + \mathbf{T}|^{-\frac{1}{2}} \\
&\times \exp\left\{ -\frac{1}{2}(\log \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{\Sigma} + \mathbf{T})^{-1}(\log \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}'(10\boldsymbol{I}_3)^{-1}\boldsymbol{\beta}) \right) \times \prod_{k=1}^{K} (\sigma_k^2)^{-3} \exp\left(-\frac{1}{\sigma_k^2} \right) \\
&\times \prod_{k=1}^{K} (\tau_k^2)^{-3} \exp\left(-\frac{1}{\tau_k^2} \right) \times \prod_{k=1}^{K} (\psi_k^2)^{-3} \exp\left(-\frac{1}{\psi_k^2} \right).
\end{aligned} \tag{3.25}
$$

3.5.2 Results

We calculated the mean posterior RMSE and MAE for each model using Equations 2.44 and 2.45, respectively. We display these values in Figure 3.9 and in Table 3.8. Model 4, the model that used two partitions and assumed local parameters within each sub-region was the best, having the lowest mean RMSE and mean MAE of all models fitted, while also having the lowest Moran's I. This suggests that Model 4 had the best predictive accuracy and accounted for spatial autocorrelation the most.

Table 3.8: Posterior means for RMSE, MAE, and Moran's I for each model fitted to the New Zealand particulate matter data.

Model	RMSE	MAE	Moran's I
1	0.902	0.754	0.342
2	1.163	0.973	0.261
3	1.207	0.998	0.279
4	0.655	0.528	0.160
5	0.745	0.598	0.182

In addition, we calculated summary statistics of the model coefficients and display these in Tables 3.9 and 3.10. In general, the estimates for the parti-

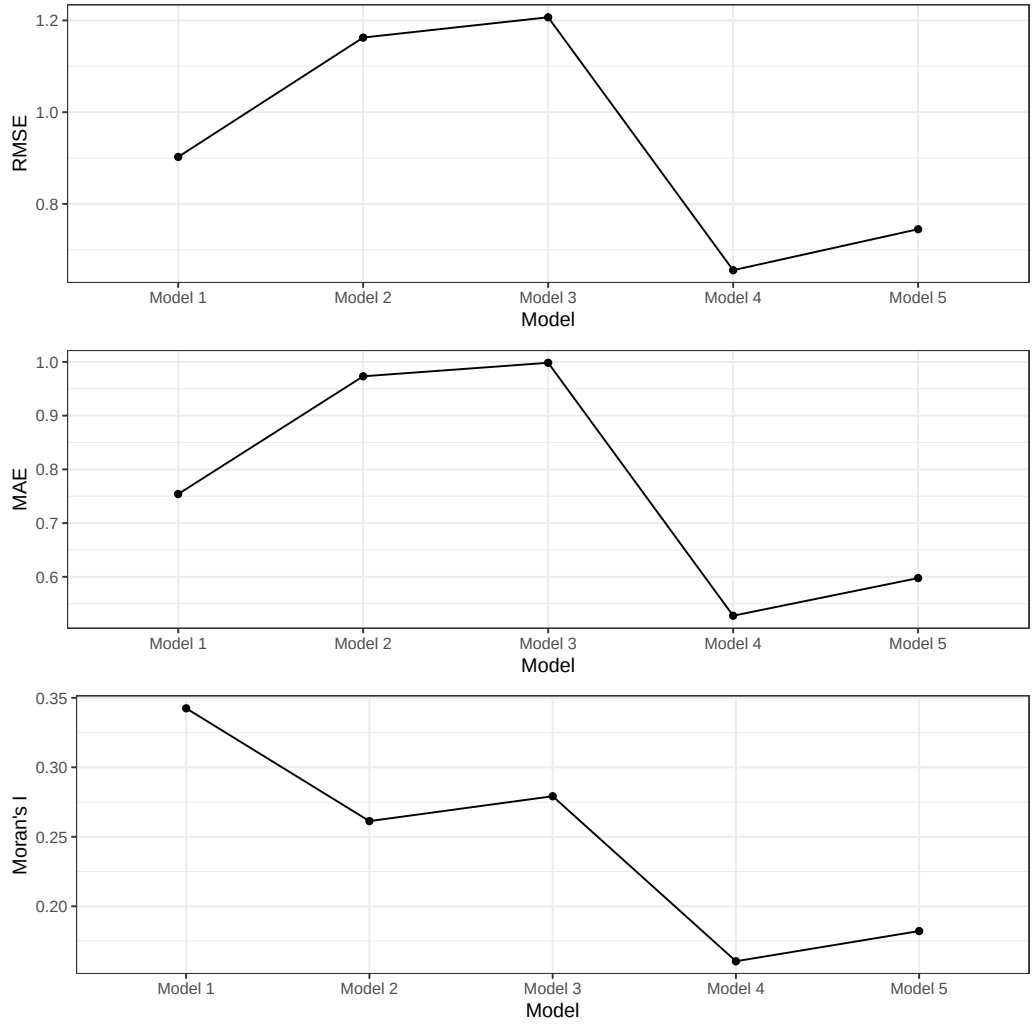


Figure 3.9: Means of the posterior distributions for RMSE, MAE, and Moran's I for each model fitted to the New Zealand particulate matter data, where each model has different degrees of partitioning. Model 1 has no partitioning, Models 2 and 4 have two partitions, and Models 3 and 5 have three partitions. Further, Models 2 and 3 assumed equal parameters across each sub-region, and Models 4 and 5 assumed local parameters within each sub-region

tioned models were more precise than that of the non-partitioned model. For example, the mean function parameters β_0, β_1, and β_2 have narrower highest posterior density (HPD) intervals for Models 2 – 5, compared to Model 1. Of all models, Model 4 has the most precise mean function parameters. This indicates that partitioning is important, and that allowing parameters between partitions to be different gives the best result.
We calculated predicted PM10 concentration using the posterior predictive distribution (Equation 2.43) of Model 4. Predictions were obtained using annual mean temperature and wind speed observations from 347 monitoring stations across New Zealand in 2013. The stations were not at the same locations that PM10 was observed. Figure 3.10 displays the interpolated surface plot of the posterior predicted annual PM10 concentration across New Zealand in 2013 using Model 4. We observed lower predicted concentrations of PM10 on the mountainous regions of the South Island and North Island. The highest predicted PM10 concentration was observed in the Canterbury region.

3.6 Conclusion

The K-means partitioned geostatistical models that we proposed provided a relatively flexible and fast way to account for non-stationarity while still allowing the use of simple stationary covariance functions. This was highlighted in the simulation studies, particularly in the spatial case. In the spatial simulation, the K-means partitioned geostatistical models (Models 2 – 9) generally provided better predictive accuracy (in terms of RMSE and MAE) when fitted to either stationary or non-stationary point referenced data.

Table 3.9: Posterior summary statistics for the mean function coefficients.

Model		Median	95% HPD Interval
Model 1	β_0	0.2163	(-0.4301, 0.8314)
	β_1	0.1581	(0.0241, 0.2563)
	β_2	-0.0504	(-0.1701, 0.0669)
Model 2	β_0	0.3764	(-0.2563, 0.9455)
	β_1	0.1667	(0.1213, 0.2124)
	β_2	-0.0009	(-0.6030, 0.6400)
Model 3	β_0	0.3056	(-0.3451, 0.9438)
	β_1	0.1724	(0.1251, 0.2223)
	β_2	0.0072	(-0.6205, 0.6359)
Model 4	β_{01}	0.0979	(-0.5096, 0.7644)
	β_{11}	0.1935	(0.0849, 0.3065)
	β_{21}	0.1545	(-0.2535, 0.5354)
	β_{02}	0.0914	(-0.5053, 0.6908)
	β_{12}	0.1750	(0.1236, 0.2372)
	β_{22}	-0.0056	(-0.1500, 0.1724)
Model 5	β_{01}	0.0241	(-0.6187, 0.6007)
	β_{11}	0.2442	(0.0797, 0.4087)
	β_{21}	0.0705	(-0.4313, 0.5955)
	β_{02}	0.0418	(-0.5648, 0.6197)
	β_{12}	0.2532	(0.1566, 0.3395)
	β_{22}	-0.1718	(-0.4078, 0.0834)
	β_{03}	0.0923	(-0.5674, 0.6617)
	β_{13}	0.1618	(0.0902, 0.2348)
	β_{23}	0.0608	(-0.2019, 0.3582)

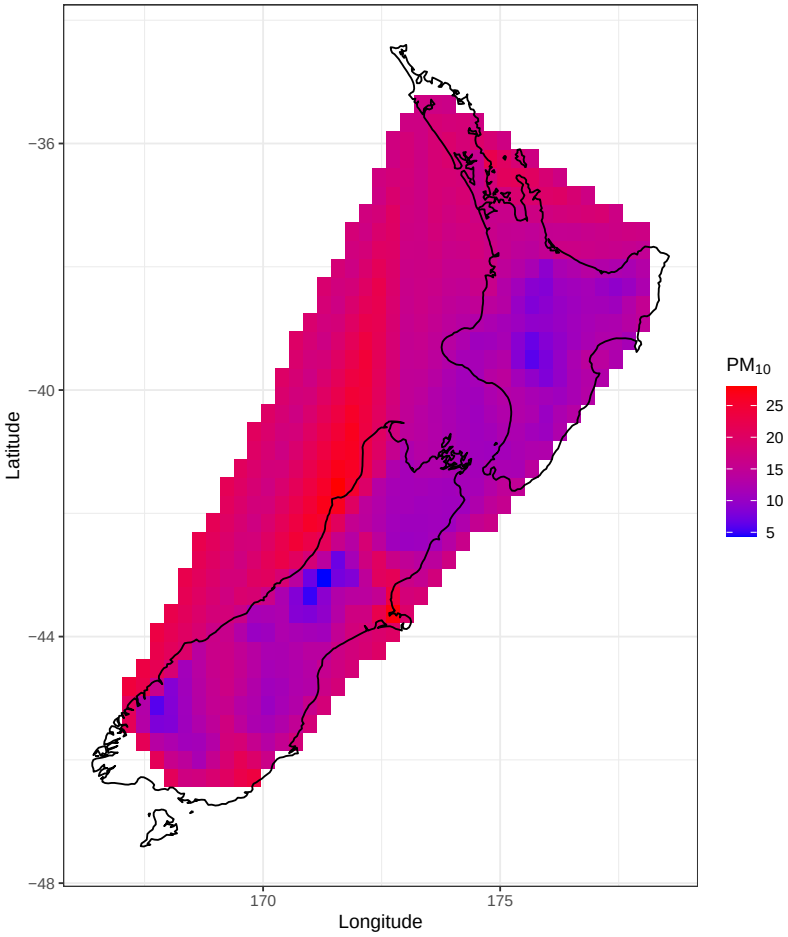


Figure 3.10: Interpolated surface plot for the posterior predicted annual particulate matter concentration in New Zealand for 2013 using Model 4.

Table 3.10: Posterior summary statistics for the covariance function parameters.

Model		Median	95% HPD Interval
Model 1	ψ	1391.1	(53.997, 22528.2)
	σ^2	0.0809	(0.0482, 0.1282)
	τ^2	0.3819	(0.0887, 1.430)
Model 2	ψ_1	0.6243	(0.0837, 4.533)
	ψ_2	0.6547	(0.0922, 5.126)
	σ_1^2	0.3044	(0.1003, 0.6575)
	σ_2^2	0.1644	(0.0756, 0.3115)
	τ_1^2	0.3022	(0.0991, 0.6932)
	τ_2^2	0.1639	(0.0741, 0.3090)
Model 3	ψ_1	0.6880	(0.0870, 48.50)
	ψ_2	0.6179	(0.0970, 3.442)
	ψ_3	0.6319	(0.1046, 3.497)
	σ_1^2	0.4341	(0.1042, 1.237)
	σ_2^2	0.2451	(0.0849, 0.5180)
	σ_3^2	0.1892	(0.0822, 0.3692)
	τ_1^2	0.4368	(0.1195, 1.239)
	τ_2^2	0.2392	(0.0879, 0.4939)
	τ_3^2	0.1925	(0.0868, 0.3934)
Model 4	ψ_1	0.5944	(0.0977, 2.588)
	ψ_2	0.6285	(0.1026, 4.490)
	σ_1^2	0.2178	(0.0869, 0.4563)
	σ_2^2	0.1616	(0.0745, 0.3050)
	τ_1^2	0.2134	(0.0882, 0.4485)
	τ_2^2	0.1611	(0.0782, 0.3057)
Model 5	ψ_1	0.5919	(0.1067, 2.832)
	ψ_2	0.6056	(0.0923, 3.189)
	ψ_3	0.5912	(0.1058, 2.834)
	σ_1^2	0.3198	(0.1029, 0.8193)
	σ_2^2	0.2249	(0.0867, 0.4838)
	σ_3^2	0.1918	(0.0874, 0.3867)
	τ_1^2	0.3291	(0.0933, 0.8471)
	τ_2^2	0.2289	(0.0848, 0.4856)
	τ_3^2	0.1909	(0.0854, 0.3850)

Chapter 4

Covariance regression network models for spatial and spatio-temporal data

The estimation of the covariance matrix $\boldsymbol{\Sigma}$ of a response vector $\boldsymbol{Y}$ is one of the key components for spatial and spatio-temporal data analysis and prediction. In Chapter 3 we showed that partitioned geostatistical models were able to handle non-stationarity by estimating stationary covariance functions for each sub-region.

When a parametric covariance function is fitted it is assumed that the covariance structure can be fully expressed using the parameters. So far we have reviewed covariance functions that depend only on the distance between observations. A more flexible approach may needed. A less restrictive approach to estimating the covariance matrix may be offered by covariance regression models.

Covariance regression involves modelling the covariance matrix of a univariate response vector as a linear combination of symmetric matrices (Liu et al., 2020; Lan et al., 2018; Zou et al., 2017). By incorporating features of the data through these symmetric matrices, a model is able to describe the covariance structure in a more flexible way, without imposing strict co-

variance functions on the data that depend only on spatial lag. A common approach is to use an adjacency matrix to take into account the network structure determined by connections between nodes of the data, which allows for the evaluation of the network effect on the covariance matrix. This approach has seen application to social networks (Liu et al., 2020; Lan et al., 2018). So far the approach has not been applied to spatial or spatio-temporal contexts.

In this chapter, we made several contributions to the covariance regression literature. First, we proposed an extension of the covariance regression network modelling framework to spatial modelling. The extension involved eliciting a strategy to estimate a network structure from the locations of an observed univariate spatial process, when the structure is unknown. We also proposed the extension to the spatio-temporal case. Both contributions were made within the Bayesian framework. Finally, we proposed a Bayesian model averaging approach (Hoeting et al., 1999) to improve predictive accuracy.

Chapter 4 begins with a concise literature review that highlights where the covariance regression approach is useful. A section is dedicated to defining the covariance regression model in a general context. This is followed by sections that introduce the spatial and spatio-temporal covariance regression network model, and the technique for estimating the network structure for both contexts. Simulation studies were performed for both the spatial and spatio-temporal cases, and we compared the covariance regression network models to a traditional Matèrn covariance function model. We followed the simulation with two case studies to illustrate the usefulness of our proposed models. We concluded the chapter with comments on the results.

4.1 Literature review

Covariance regression is the name given to a class of models that employ a regression framework to model the covariance matrix $\boldsymbol{\Sigma}$ of a response vector $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ as a linear combination of known symmetric matrices. The model for $\boldsymbol{Y}$ takes the form,

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{4.1}$$

where $\boldsymbol{\mu}$ is the mean vector and is modelled in the typical way as a linear combination of parameters, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$, and p covariates, $\boldsymbol{X} = (\boldsymbol{1x}_1, ..., \boldsymbol{x}_p)$,

$$\boldsymbol{\mu} = \boldsymbol{X\beta}. \tag{4.2}$$

The covariance matrix takes the form of a linear combination of parameters, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_r)'$, and r known symmetric matrices, $\boldsymbol{B}_0, ..., \boldsymbol{B}_r$,

$$\boldsymbol{\Sigma} = \sum_{k=0}^{r} \gamma_k \boldsymbol{B}_k. \tag{4.3}$$

The matrices $\boldsymbol{B}_0, ..., \boldsymbol{B}_r$ are assumed to be linearly independent, and it is assumed that at least one set of parameters $\boldsymbol{\gamma}$ results in $\boldsymbol{\Sigma}$ positive definite. Covariance regression was described in Anderson et al. (1973), where estimation procedures were detailed for several cases of the mean vector and covariance matrix being known or unknown. The estimation procedure for the covariance matrix involved maximum likelihood estimation. In addition, asymptotic efficiency of the estimators was discussed. Later, in Szatrowski (1980) and Zwiernik et al. (2014), properties of the covariance estimates under the linear structure were examined.

In many applications, covariates have been shown to not only affect the mean vector, but also play an important role when determining the covariance matrix of a process (Schmidt et al., 2011). The covariates that affect the covariance matrix may not necessarily be the same as the covariates that contribute to the linear combination for the mean vector. As

such, we defined the r covariates assumed to have an effect on the covariance matrix as $\boldsymbol{w}_1, ..., \boldsymbol{w}_r$, where $\boldsymbol{w}_k = (w_{k1}, ..., w_{kn})'$ is an n-dimensional observed covariate vector with each w_{ki} corresponding to an observed response y_i, for $k = 1, ..., r$. Covariance regression was the focus of Zou et al. (2017), in which, the explicit regression relationship between the covariance matrix $\boldsymbol{\Sigma}$ and the covariates, $\boldsymbol{w}_k$ was explored. In the article, Zou et al. (2017) proposed a methodology that allowed the covariate information to be represented as symmetric matrices and explicitly linked to the covariance matrix through Equation 4.3. The methodology comes from adapting the concept of pairwise comparisons (Johnson & Wichern, 1992), which considers measures of similarity or distance of covariate values between pairs of subjects i to build the symmetric matrices. In other words,

$$\boldsymbol{\Sigma} = \sum_{k=0}^{r} \gamma_k \boldsymbol{B}_k = \sum_{k=0}^{r} \gamma_k W(\boldsymbol{w}_k), \tag{4.4}$$

where $W(\boldsymbol{w}_k) = (\delta(w_{ki}, w_{kj}))_{n \times n}$ is a matrix with elements that measure the similarity (or distance), δ, of covariate $\boldsymbol{w}_k$ between each pair of subjects i and j.

A particular measure of similarity between covariates that was investigated in Lan et al. (2018) was the adjacency matrix. This was motivated by the practical example of a mobile network. In that case, the response variable was the logarithm of monthly call duration measured in log minutes, and they proposed a regression framework to model the covariance matrix as a function of the adjacency matrix of the mobile network. This leads to the so called covariance regression network (CRN) framework.

4.2 Covariance regression network model

We now explicitly introduce the covariance regression network (CRN) model proposed in Lan et al. (2018). Consider a network of nodes, $i = 1, ..., n$, and

let $\boldsymbol{A} = (a_{ij})_{n\times n}$ denote the adjacency matrix. We define,

$$a_{ij} = \begin{cases} 1 \text{ if } i,j \text{ connected,} \\ 0 \text{ if } i,j \text{ not connected and } i \neq j, \\ 0 \text{ if } i \neq j. \end{cases} \tag{4.5}$$

The adjacency matrix $\boldsymbol{A}$ identifies the pairwise connections that exist between nodes. This gives the structure of the network. Let also $\boldsymbol{A}^k = (a_{ij}^{(k)})_{n\times n}$ be the matrix where $a_{ij}^{(k)}$ represents the number of paths of length k from node i to node j. That is to say $\boldsymbol{A}^k$ identifies all pairwise connections separated by paths of length k. For completeness, we define $\boldsymbol{A}^0 = \boldsymbol{I}_n$, the n-dimensional identity matrix.
We provided an example of a network of five nodes in Figure 4.1. In this example, we define the matrices, $\boldsymbol{A}$, $\boldsymbol{A}^2$, and $\boldsymbol{A}^3$ as,

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{A}^2 = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 3 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}, \boldsymbol{A}^3 = \begin{bmatrix} 0 & 3 & 1 & 1 & 1 \\ 3 & 2 & 4 & 5 & 1 \\ 1 & 4 & 2 & 4 & 1 \\ 1 & 5 & 4 & 3 & 3 \\ 1 & 1 & 1 & 3 & 0 \end{bmatrix}.$$

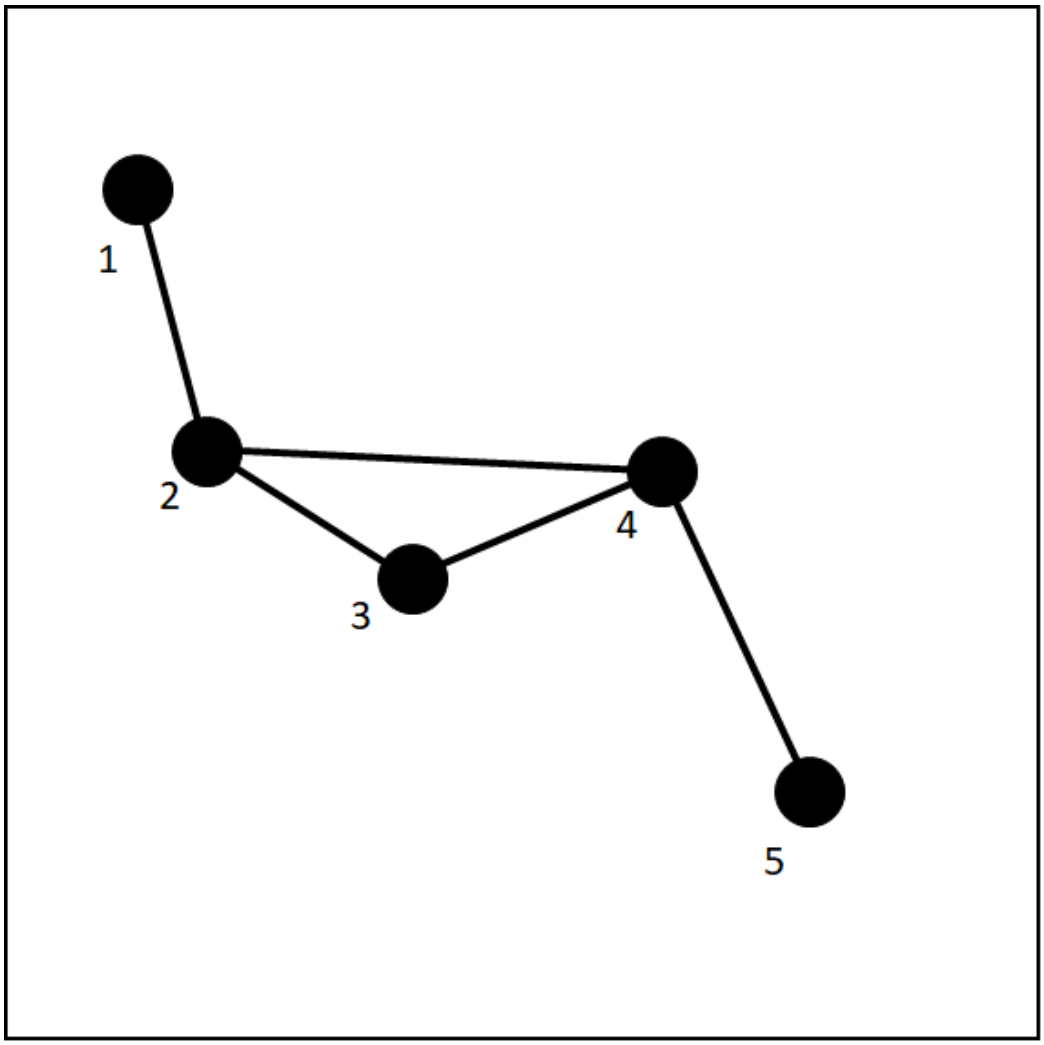
We now introduce the model that connects the network structure information to the covariance matrix. Let $\boldsymbol{y} = (y_1, ..., y_n)'$ be a univariate response vector, where y_i is an observed response associated with node i. The model for $\boldsymbol{y}$ is simply,

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{4.6}$$

where $\boldsymbol{\mu}$ is the mean vector and can be modelled in the typical way, as a linear combination of covariates and parameters, $\boldsymbol{X\beta}$. Here, $\boldsymbol{\Sigma}$ is the covariance matrix, and is modelled as a linear combination of parameters, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_r)'$, and adjacency matrices (Lan et al., 2018),

$$\boldsymbol{\Sigma} = \sum_{k=0}^{r} \gamma_k \boldsymbol{A}^k. \tag{4.7}$$

Figure 4.1: Network with five connected nodes.



The parameters $\boldsymbol{\gamma}$ describe the strength of influence that the connected points have on the covariance. For example, a larger value for γ_3 implies that the contribution of $\boldsymbol{A}^3$ on the covariance matrix is bigger, and suggests that points that are implicitly connected via paths of length 3 have an important effect on the covariance. The model assumes that the network structure is known or correctly specified. We now propose our extension of the CRN model to spatial data.

4.3 Spatial covariance regression network model for point reference data

In the spatial setting, we replaced the network of n nodes in Section 4.2 with a network of n fixed locations, denoted by $i = 1, ..., n$. The CRN model described by Equations 4.6 and 4.7 require that the network structure is known. However, we typically do not know the network structure of spatial data, which is to say we do not know how the observations at each location are connected. For spatial data, we propose the definition of connection: *any two point locations separated by a distance that is less than a bandwidth parameter, d, are connected*. We propose that the network structure can be approximated using the location data, by estimating the adjacency matrix.

The network structure of a set of n fixed locations can be approximated by estimating an adjacency matrix, $\boldsymbol{A} = (a_{ij})_{n\times n}$, using a distance function. There are many different distance functions to choose from and in this thesis, we estimate the adjacency matrix with the distance function,

$$a_{ij} = \begin{cases} 1 \text{ if } d_{ij} < d \\ 0 \text{ otherwise,} \end{cases} \tag{4.8}$$

and $a_{ii} = 0$, where d_{ij} is the Euclidean distance between locations i and j, and d is the bandwidth parameter. We propose that d can be determined

in a similar way to how the range parameter of a variogram is chosen. For instance, d can be chosen as the threshold distance that the researcher believes above which pairs of points are likely to be uncorrelated. Therefore, the approximated network structure is based on Tobler's law, in that locations that are closer to each other are more likely to be connected than those further apart.

In order to remove some of the subjectivity surrounding the determination of the bandwidth, d, we propose using Bayesian model averaging (Hoeting et al., 1999) over a range of CRN models the network structure is approximated by adjacency matrices using different bandwidths in Equation 4.8. In doing so, we average over many candidate models, which removes some of the uncertainty around choosing a correctly specified network structure.

4.3.1 Bayesian model averaging

In this thesis, we followed the Bayesian model averaging procedure of Fragoso et al. (2018). Let each candidate model in consideration be denoted by M_h, for $h = 1, ..., H$, which represent a set of probability distributions encompassing the likelihood function $f(\boldsymbol{Y}|\boldsymbol{\theta}_h, M_h)$ of the observed data $\boldsymbol{Y}$ in terms of model specific parameter vector $\boldsymbol{\theta}_h$ and a set of prior probability densities, $\pi(\boldsymbol{\theta}_h|M_h)$. Recall that given a model under the Bayesian framework, we obtain the posterior distribution using Bayes' theorem, and this is given by Equation 2.34 in Section 2. The denominator of the posterior distribution is referred to as the model's marginal likelihood or model evidence, denoted by,

$$\pi(\boldsymbol{Y}|M_h) = \int f(\boldsymbol{Y}|\boldsymbol{\theta}_h, M_h)\pi(\boldsymbol{\theta}_h|M_h)d\boldsymbol{\theta}_h. \tag{4.9}$$

Bayesian model averaging then adds another layer by assuming a prior distribution, $\pi(M_h)$ over the set of considered candidate models describing the prior uncertainty over each model's capability to accurately describe

the data. The posterior model probabilities given the observed data are given by,

$$\pi(M_h|\boldsymbol{Y}) = \frac{\pi(\boldsymbol{Y}|M_h)\pi(M_h)}{\sum_{m=1}^{H}\pi(\boldsymbol{Y}|M_m)\pi(M_m)}, \tag{4.10}$$

which represents the support for each considered candidate model by the observed data. Using the posterior model probabilities, we can construct a marginal posterior distribution for the predicted value, $\hat{\boldsymbol{Y}}$ across all considered models,

$$\pi(\hat{\boldsymbol{Y}}|\boldsymbol{Y}) = \sum_{h=1}^{H}\pi(\hat{\boldsymbol{Y}}|\boldsymbol{Y}, M_h)\pi(M_h|\boldsymbol{Y}), \tag{4.11}$$

which is an average of all posterior distributions weighted by each posterior model probability.

4.4 Spatio-temporal covariance regression network model for point reference data

We propose that a CRN model can be fitted to point reference data in a spatio-temporal context. In the spatio-temporal setting, we replaced the network of n nodes in Section 4.2 with a network of n fixed locations, denoted by $i = 1, ..., n$. At each location, a response is observed at T time points. For the purpose of this thesis, we assume the locations remain constant over time. The model for $\boldsymbol{y}_t$ is given by,

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t), \tag{4.12}$$

where $\boldsymbol{\mu}_t$ is the mean vector and can be modelled in the typical way, as a linear combination of covariates and parameters, $\boldsymbol{X}_t\boldsymbol{\beta}$. Here, $\boldsymbol{\Sigma}_t$ is the covariance matrix, and is modelled as a linear combination of parameters, $\boldsymbol{\gamma}_t = (\gamma_{1t}, ..., \gamma_{rt})'$, and adjacency matrices,

$$\boldsymbol{\Sigma}_t = \sum_{k=0}^{r}\gamma_{kt}\boldsymbol{A}^k, \quad \text{for } t = 1, ..., T. \tag{4.13}$$

We assumed the network structure does not change over time.
The CRN model described by Equations 4.12 and 4.13 requires that the network structure is known or correctly specified. However, we typically do not know the network structure of spatial data, which is to say we do not know how the observations at each location are connected. We approximate the network structure in the same way as the spatial case by estimating the adjacency matrix, using Equation 4.8, and determining d by prior knowledge or fitting a range of models using different values of d and performing Bayesian model averaging.
We now perform several simulation experiments to evaluate the predictive accuracy of the CRN model on simulated spatial and spatio-temporal data.

4.5 Simulation

We evaluated the performance of CRN models within a Bayesian framework on simulated spatial and spatio-temporal data. We fitted several models that were determined by assuming the network structure was unknown and estimated using an adjacency matrix determined by Equation 4.8. The aim of the simulation experiment was to investigate the performance of CRN models in terms of ability to accurately predict, and account for autocorrelation, in the values of a dependent variable. Furthermore, we compared these abilities to a traditional Matèrn covariance model and a CRN model defined by a correctly specified network structure. For the comparisons, we computed three measures of accuracy and a measure of residual spatial autocorrelation over the set of models. The measures of accuracy that were used are the root mean square error (RMSE, Equation 2.44), the mean absolute error (MAE, Equation 2.45), and the average estimation error for the covariance matrix (COV, Equation 2.46). The measure of residual spatial autocorrelation that was used is Moran's I (Equation 2.8), calculated on the residuals. We performed the simulation experiments separately for the spatial case and the spatio-temporal case.

4.5.1 Spatial simulation

A simulation experiment was conducted to evaluate the performance of CRN models with estimated network structures on simulated data with a spatial structure.
We randomly generated $p = 200$ longitude (s_{long}) and latitude (s_{lat}) values from a unit square,

$$s_{\text{long}} = \text{U}(0, 1),$$
$$s_{\text{lat}} = \text{U}(0, 1).$$

We then simulated a network structure taking in to account Tobler's Law (Tobler, 1970), by connecting each pair of locations that were separated by a Euclidean distance of bandwidth d or smaller. The value of d was chosen such that the resulting network structure had a network density of 5%, where, we define network density, ND as,

$$\text{ND} = \frac{\text{Number of connections}}{n(n-1)/2}. \tag{4.14}$$

In this way, the bandwidth was calculated to be $d = 0.2023$. The network structure was represented by an adjacency matrix given by,

$$\boldsymbol{A} = (a_{ij})_{200\times200}, \tag{4.15}$$

where a_{ij} was calculated using Equation 4.8. A dependent variable, $\boldsymbol{y}$, was then simulated from the model,

$$\boldsymbol{y} = \beta_0\mathbf{1} + \boldsymbol{\varepsilon}, \tag{4.16}$$

where $\mathbf{1}$ is a 200×1 vector of 1's, β_0 is the intercept, and $\boldsymbol{\varepsilon}$ are the errors for the spatial process that introduced spatial autocorrelation to $\boldsymbol{y}$. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y} \sim \text{N}(\beta_0\mathbf{1}, \boldsymbol{\Sigma}), \tag{4.17}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix modelled by the CRN model,

$$\boldsymbol{\Sigma} = \gamma_0 \boldsymbol{I}_{200} + \gamma_1 \boldsymbol{A} + \gamma_2 \boldsymbol{A}^2. \tag{4.18}$$

Here, γ_0 represents the variance of the measurement errors, while $\gamma_1 \boldsymbol{A} + \gamma_2 \boldsymbol{A}^2$ represents the purely spatial covariance matrix. The model assumed that only connections between locations of path lengths 1 and 2 contribute to the covariance matrix. This was inline with the simulation studies of Zou et al. (2017), Lan et al. (2018), and Liu et al. (2020). The parameters β_0, γ_0, γ_1, and γ_2 were chosen specifically to focus on estimation of the covariance matrix parameters. We set $\beta_0 = 0$, $\gamma_0 = 1$, $\gamma_1 = 0.5$, and $\gamma_2 = 0.2$, inline with Lan et al. (2018) who provided a demonstrated example. The parameters were chosen such that $\gamma_0 > \gamma_1 > \gamma_2$, which produced a covariance matrix that was positive definite. The data were sampled using Cholesky factorization for a Gaussian process (Algorithm 2, Rue & Held (2005)). We generated 30 sets of simulated data and fitted the models to each set in order to stabilize the variation due to generating the data.

Figure 4.2 displays the interpolated surface plot of $\boldsymbol{y}$ for one set of simulated data. It provided evidence for the presence of spatial autocorrelation within the dependent variable. We observed clusters of higher values at the lower right region and upper region, as well as clusters of lower values observed at the upper left and lower right regions of the plot. Moran's I was calculated to confirm the presence of spatial autoccorelation in the simulated y values. We calculated $I = 0.13$, with a p=value for the two-sided test for presence of spatial autocorrelation less than 2.2×10^{-16}, confirming the presence of significant spatial autocorrelation within the dependent variable. Similar findings were obtained for the 29 other sets of simulated data.

We fitted several CRN models to the simulated spatial data sets. We fitted Model 1, a CRN model that used the same known network structure used to simulate the data, where the bandwidth d was chosen such that the resulting network had a density of 5%. Models 2 – 10, fitted next, were

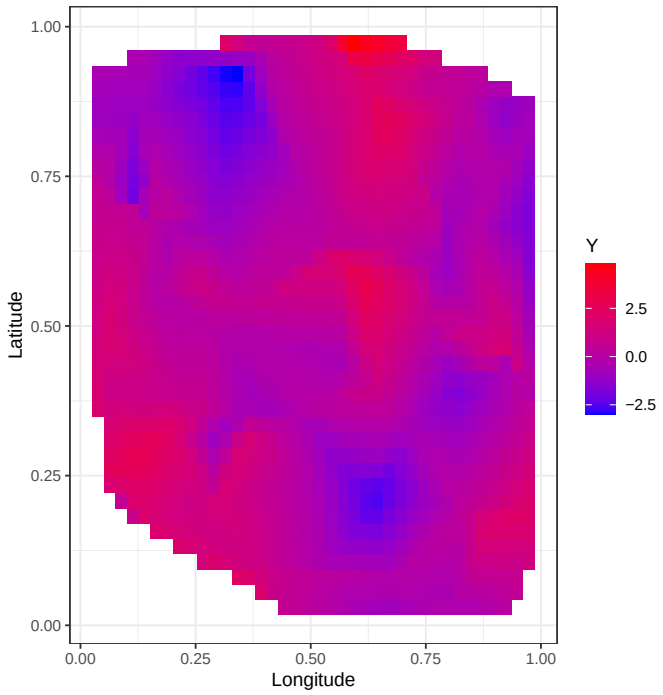


Figure 4.2: Interpolated surface plots of the simulated dependent variable. Spatial autocorrelation is exhibited for $\boldsymbol{y}$. There are clusters of higher values near the bottom left, center, and top right, and clusters of lower values displayed at the top left, bottom, and far right.

a set of CRN models that assumed the network structure of the data was unknown and estimated by the adjacency matrix using Equation 4.8. The models were determined by changing the bandwidth, d, such that the approximating networks had densities of 1% – 4% and 6% – 10% for Models 2 – 10, respectively. For each model, the covariance matrix was estimated using Equation 4.18. In addition to fitting CRN models, we also fitted Model 11, a traditional exponential Matèrn covariance model, defined by Equation 2.31 from Chapter 2. For each model, we produced posterior distributions of predicted values, $\hat{\boldsymbol{y}}$, which were used to calculate the measures of predicted accuracy and residual spatial autocorrelation. We also produced two further posterior distributions of predicted values by Bayesian model averaging over Models 1 – 10 and over Models 2 – 10, using Equations 4.10 and 4.11. For simplicity, we denoted these sets of predicted values BMA 1 and BMA 2. The set BMA 1 made intuitive sense since it represented the posterior distribution of predicted values weighted by the posterior model probability for each candidate CRN model fitted to the data. However, it was likely that Model 1 would perform well because the network structure was correctly specified. This would result in a large posterior model probability for Model 1 and would lend to a large contribution from Model 1 to the set of predicted values. However, in reality, it is unlikely that the true network structure will be known or easily defined. Therefore, we produced a second set of Bayesian model averaged predicted values, BMA 2, averaged over the models where the network structure was misspecified. A description of each model and set of predicted values is given in Table 4.1. We described each model in more detail in the following sections.

Table 4.1: Description of Models 1 – 11, BMA 1, and BMA 2 used to obtain the 13 sets of predictions.

Prediction set	Method	Network structure	Network density
1	Model 1	Correctly specified	5%
2	Model 2	Misspecified	1%
3	Model 3	Misspecified	2%
4	Model 4	Misspecified	3%
5	Model 5	Misspecified	4%
6	Model 6	Misspecified	6%
7	Model 7	Misspecified	7%
8	Model 8	Misspecified	8%
9	Model 9	Misspecified	9%
10	Model 10	Misspecified	10%
Prediction set	Method	Covariance structure	
11	Model 11	Exponential	
Prediction set	Method	Averaged over	
12	BMA 1	Models 1 – 10	
13	BMA 2	Models 2 – 10	

Model 1

We first fitted Model 1, a CRN model (Lan et al., 2018; Liu et al., 2020) where the network structure of the data was assumed to be known and correctly specified. The network structure of the data was the same as the one used to generate the simulated data. The bandwidth d was such that the network had a density of 5%. The model is described by Equations 4.16 – 4.18. The data likelihood is given as,

$$f(\boldsymbol{y}|\beta_0, \boldsymbol{\gamma}, \boldsymbol{A}) = (2\pi)^{-\frac{200}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \beta_0 \mathbf{1})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \beta_0 \mathbf{1}) \right\}, \quad (4.19)$$

where $\boldsymbol{\Sigma} = \gamma_0 \boldsymbol{I}_{200} + \gamma_1 \boldsymbol{A} + \gamma_2 \boldsymbol{A}^2$, and $\boldsymbol{A}$ is same adjacency matrix used to simulate the data, defined by Equation 4.15, where $d = 0.2023$.

Model 2 – 10

Models 2 – 10 were fitted next and are CRN models (Lan et al., 2018; Liu et al., 2020) where the network structure of the data was assumed to be unknown. We estimated the network structure using the adjacency matrix, determined by changing the bandwidth d, such that the approximating networks had densities of 1% – 4% and 6% – 10% for Models 2 – 10, respectively. The models are described by Equations 4.16 – 4.18 and the data likelihoods are given by,

$$f(\boldsymbol{y}|\beta_0, \boldsymbol{\gamma}, \boldsymbol{A}) = (2\pi)^{-\frac{200}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \beta_0 \mathbf{1})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \beta_0 \mathbf{1}) \right\}, \quad (4.20)$$

where $\boldsymbol{\Sigma} = \beta_0 \boldsymbol{I}_{200} + \beta_1 \boldsymbol{A} + \beta_2 \boldsymbol{A}^2$.

Model 11

Model 11 was then fitted and is a traditional Matèrn covariance model with smoothness parameter $\nu = 0.5$, corresponding to the exponential covariance model. The measurement equation for $\boldsymbol{y}$ under Model 11 is given by Equation 2.19, where the covariance matrix is $\boldsymbol{C} = \boldsymbol{\Sigma} + \mathrm{T}$, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{R}$ and $\mathrm{T} = \tau^2 \boldsymbol{I}_{200}$. Model 11 assumed a covariance function for the spatial process, $\boldsymbol{\Sigma}$, that depended on pairwise distances between locations. The function is defined by,

$$\boldsymbol{\Sigma} = \sigma^2 \exp(-\boldsymbol{D}/\psi), \quad (4.21)$$

where $\boldsymbol{D}$ is a 200×200 matrix of pairwise distances between locations, σ^2 is the spatial process variance, and ψ is the spatial correlation strength parameter that measures the strength of correlation between two locations.

The data likelihood is given by,

$$f(\boldsymbol{y}|\beta_0, \sigma^2\tau^2, \psi, \boldsymbol{D}) = (2\pi)^{-\frac{200}{2}}|\boldsymbol{C}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \beta_0\mathbf{1})'\boldsymbol{C}^{-1}(\boldsymbol{y} - \beta_0\mathbf{1})\right\}. \tag{4.22}$$

BMA 1 and BMA 2

BMA 1 and BMA 2 represent the posterior distributions of predicted values that were calculated using Bayesian model averaging over Models 1 – 10 and over Models 2 – 10, respectively. For each set, we used the R package `bridgesampling`, which calculates the model evidence for each Model using Equation 4.9. We then compute the posterior model probabilities for Models 1 – 10 and Models 2 – 10, separately, using Equation 4.10. The posterior model probabilities were used to calculate the weighted average posterior predictions, given by Equation 4.11.
We used MCMC to fit the models to both simulated sets of data and for each repetition. For each model we assigned a vague prior to the intercept β_0,

$$\beta_0 \sim \mathrm{N}(0, 1000).$$

CRN models have not been estimated using the Bayesian framework so far. For Models 1 – 10, it was essential that the posterior distributions for $\boldsymbol{\gamma}$ were such that the covariance matrix was positive definite. We believed that the choice of prior distribution for $\boldsymbol{\gamma}$ could have large implications on the positive definiteness of the estimated covariance matrix. We explored combinations of values for $\boldsymbol{\gamma}$ that would result in a positive definite covariance matrix. We chose to assign the following uniform priors to $\boldsymbol{\gamma}$ to minimise the number of resulting posterior covariance matrices that would not be positive definite:

$$\gamma_0 \sim \mathrm{U}(0, 5),$$
$$\gamma_1 \sim \mathrm{U}(0, 2),$$
$$\gamma_2 \sim \mathrm{U}(0, 1).$$

The posterior distribution for Models 1 – 10 is,

$$f(\beta_0, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{A}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \beta_0\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \beta_0\mathbf{1})\right\} \exp\left\{ -\frac{1}{2}\frac{\beta_0^2}{1000}\right\}, \tag{4.23}$$

where $\boldsymbol{\Sigma} = \gamma_0\boldsymbol{I}_{200} + \gamma_1\boldsymbol{A} + \gamma_2\boldsymbol{A}^2$.
For Model 11, we assigned vague priors to the covariance parameters, σ^2, τ^2, and ψ,

$$\begin{aligned}\sigma^2 &\sim \text{IG}(2,1),\\ \tau^2 &\sim \text{IG}(2,1),\\ \psi &\sim \text{IG}(2,1).\end{aligned}$$

The posterior distribution for Model 11 is,

$$\begin{aligned} f(\beta_0, \sigma^2, \tau^2, \psi|\boldsymbol{y}, \boldsymbol{D}) \propto |\boldsymbol{C}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \beta_0\mathbf{1})'\boldsymbol{C}^{-1}(\boldsymbol{y} - \beta_0\mathbf{1})\right\} \\ \exp\left\{ -\frac{1}{2}\frac{\beta_0^2}{1000}\right\}(\sigma^2)^{-3}\exp\left(-\frac{1}{\sigma^2}\right)(\tau^2)^{-3}\exp\left(-\frac{1}{\tau^2}\right)(\psi)^{-3}\exp\left(-\frac{1}{\psi}\right), \end{aligned} \tag{4.24}$$

where $\boldsymbol{C} = \boldsymbol{\Sigma} + \text{T} = \sigma^2\exp(-\boldsymbol{D}/\psi) + \tau^2\boldsymbol{I}_{200}$.
For each repetition of the simulation, each model was run for two chains, each with 7500 iterations. The chains converged to stationary distributions slowly and so 6750 (90%) of the iterations were discarded as warm-up. We thinned each chain by 2, to minimize autocorrelation in the posterior samples affording posterior draws of size 750. Trace plots, density curves, and autocorrelation plots were checked to determine that the posterior samples converged to stationary distributions. For conciseness, diagnostic plots were given for Model 1 only, in Figure B.1. In addition, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, given in Table B.1. Furthermore, the convergence diagnostics were reported for the first repetition of the simulation only. We noted similar convergence for each repetition. For each model fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary dis-

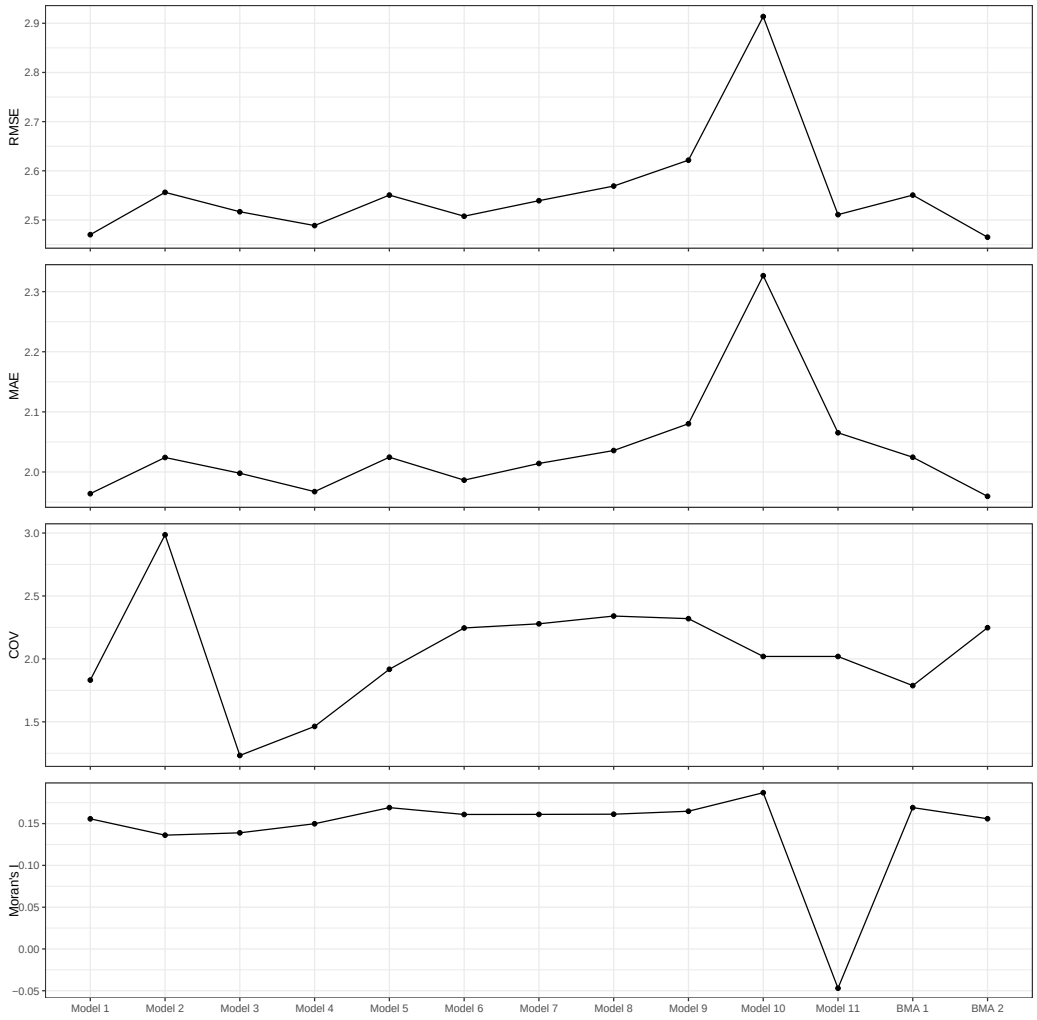


Figure 4.3: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I, summarised over the 30 sets of simulated spatial data.

tributions, and appropriate exploration and mixing of the posterior distributions. Furthermore, $\hat{R}$ was found to be close to 1 for each parameter, indicating convergence (Brooks & Gelman, 1998).

Figure 4.3 displays the medians of the posterior distributions for RMSE (Equation 2.44), MAE (Equation 2.45), and Moran's I (Equation 2.8), summarised by the median over the 30 sets of simulated data. In addition, the

average estimated error in log scale for the covariance matrix, (COV, Equation 2.46) is displayed for each model. Table 4.2 collects these measures.

Table 4.2: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I, summarised over the 30 sets of simulated spatial data.

Method	Predictive accuracy			Autocorrelation
for prediction	RMSE	MAE	COV	Moran's I
Model 1	2.470	1.964	1.832	0.156
Model 2	2.556	2.024	2.986	0.136
Model 3	2.517	1.998	1.233	0.139
Model 4	2.489	1.967	1.463	0.150
Model 5	2.551	2.025	1.918	0.169
Model 6	2.508	1.986	2.246	0.161
Model 7	2.539	2.014	2.279	0.161
Model 8	2.569	2.036	2.340	0.161
Model 9	2.622	2.080	2.320	0.165
Model 10	2.914	2.327	2.019	0.187
Model 11	2.511	2.065	2.019	-0.047
BMA 1	2.551	2.025	1.788	0.169
BMA 2	2.465	1.960	2.248	0.156

The CRN model with the lowest posterior median RMSE and posterior median MAE was Model 1, averaged over the 30 repetitions of simulated spatial data. This suggests that Model 1, a model that correctly specified the covariance matrix provided the best predictive accuracy. This result was not surprising, since we expected Model 1, the data generating model, to outperform each model.

When we assumed that the network structure was unknown, we estimated it using the adjacency matrix and a bandwidth, d determined by network density. Models 2 – 10 therefore represented CRN models where the network structure was estimated with increasing connectivity. We found that for Models 2 – 10, where the network structures were estimated with densities 1% – 4%, and, 6% – 10% there was an increasing trend in posterior median RMSE and MAE, with the model that resulted in the

highest posterior median RMSE and MAE being Model 10. This suggests that increasing the number of connections of path lengths 1 and 2 reduces the predictive accuracy.

When we compared the posterior median RMSE and MAE of Models 1 – 10 with that of the Matèrn covariance model (Model 11), we saw that Model 1 provided better predictive accuracy than the traditional model. Further, Models 4 and 6 had comparable posterior median RMSE and MAE to that of Model 11. This shows that covariance regression network modelling is a viable methodology for predicting spatial data.

When we took the average predicted values from Models 1 – 10 weighted by each models posterior model probability (BMA 1), we found the posterior median RMSE and MAE were higher than that of Models 3 – 6. Model 1 was found to have the largest posterior model probability and so Model 1 contributed the most to BMA 1. This suggests that averaging over a range of CRN models with different estimated network structures (including the true network structure), we obtain similar predictive accuracy compared to the individual CRN models but worse predictive accuracy than that of the Matèrn model. Averaging over only those CRN models where the true network structure was not included (BMA 2), we found the lowest values for posterior median RMSE and MAE, compared to all other CRN models and BMA 1. This suggests that CRN models have potential to be used to make accurate predictions for spatial point reference data.

Moran's I was calculated for each set of posterior residuals for each model, using Equation 2.47 and the posterior median Moran's I across each set of simulated data was plotted in Figure 4.3. We found that each model was not able to account for spatial autocorrelation within the simulated data. For each model, the posterior median of Moran's I the same or larger than Moran's I for the simulated data, which was $M = 0.13$. Between Models 1 – 10, we found a generally increasing trend in posterior median Moran's I. This suggests that as the estimated network structures connectivity increased, the amount of residual spatial autocorrelation increased.

This seemed counter-intuitive, perhaps pertaining to variation among the simulate data sets. Both BMA 1 and BMA 2 produced relatively similar Moran's I values to that of Models 6 and 5, respectively. This suggests that when we average over a range of CRN models with different estimated network structures (including the known network structure) residual spatial autocorrelation is averaged over that of the models, but not reduced over all. Model 11, the Matèrn covariance model gave a Moran's I value of $I = -0.0773$. While this value is closer to 0 than that of Models 1 – 10, and BMA 1 and 2, it is negative. This is due to incorrectly specifying the covariance matrix.

Another measure of accuracy is that given by the average estimated error (COV) between the model covariance matrix and the true covariance matrix. The posterior median COV is displayed in Figure 4.3 for each model. Of the ten individual CRN models, Models 1 – 10, Model 2 had the lowest posterior median COV, which suggests that the posterior distribution of covariance matrices for Model 2 were the closest in Frobenius Norm to the true covariance matrix. As the estimated network structures became more connected in terms of network density, we observed a general increasing trend in the posterior median COV. When we weighted the posterior distributions of the covariance matrix from Models 1 – 10, we found a similar posterior median COV for BMA 1 compared to Model 5. This was expected, since Model 1 had the largest weight in the model averaged BMA 1. The posterior median COV BMA 2 was lowere than that of BMA 1, but higher than that of Model 1. This indicated that the model averaging was able to produce posterior distributions for the covariance matrix close to the true covariance matrix.

4.5.2 Spatio-temporal simulation

In addition to the spatial simulation, a simulation experiment was conducted to evaluate the performance of the CRN models with estimated

network structures on simulated data with a spatio-temporal structure.
We used the same randomly generated $n = 200$ longitude (s_{long}) and latitude (s_{lat}) values from the spatial simulation study. The network structure for the data was also the same. The network structure was represented by an adjacency matrix given by Equation 4.15 where a_{ij} was calculated using Equation 4.8. We assumed that the network structure for the spatio-temporal data remained constant across time.
A dependent variable, $\boldsymbol{y}_t$, was then simulated for $T = 2$ time points, from the model,

$$\boldsymbol{y}_t = \beta_{0t}\mathbf{1} + \boldsymbol{\varepsilon}_t, \tag{4.25}$$

where $\mathbf{1}$ is a 200×1 vector of 1's, β_{0t} is the temporally varying mean at time t, and $\boldsymbol{\varepsilon}_t$ are the errors at time t for the spatio-temporal process, and $t = 1, 2$. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y}_t \sim \mathrm{N}(\beta_{0t}\mathbf{1}, \boldsymbol{\Sigma}_t), \tag{4.26}$$

where $\boldsymbol{\Sigma}_t$ is the covariance matrix at time t. We assumed that the covariance matrix was dependent on time to induce spatio-temporal autocorrelation. The covariance matrix at time t was modelled by the CRN model,

$$\boldsymbol{\Sigma}_t = \gamma_{0t}\boldsymbol{I}_{200} + \gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2. \tag{4.27}$$

Here, γ_{0t} represents the measurement variance at time t, while $\gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2$ represents the spatio-temporal covariance matrix at time t (Lan et al., 2018). Like the spatial simulation, the model assumed that only connections between locations of path lengths 1 and 2 contributed to the covariance matrix. This was inline with the spatial simulation studies of Zou et al. (2017), Lan et al. (2018), and Liu et al. (2020). The parameters β_{01}, β_{02}, γ_{01}, γ_{02}, γ_{11}, γ_{12}, γ_{21}, and γ_{22} were chosen specifically to focus on estimation of the covariance matrix parameters. In order to induce a temporal trend in the dependent variable, we chose different values for $\beta 0t$ for each t. We set $\beta_{01} = 1$ and $\beta_{02} = 0.5$. In addition, we induced spatio-temporal autocorrelation by choosing different values for γ_{0t}, γ_{1t}, and γ_{2t}, for each t.

We set $\gamma_{01} = 2$, $\gamma_{11} = 1$, $\gamma_{21} = 0.5$, in line with the demonstrated example in Lan et al. (2018) and arbitrarily set $\gamma_{02} = 1$, $\gamma_{12} = 0.5$, and $\gamma_{22} = 0.25$. The covariance parameters were chosen such that $\gamma_{0t} > \gamma_{1t} > \gamma_{2t}$, which produced a covariance matrix that was positive definite. The data was sampled using Cholesky factorization for a Gaussian process (Algorithm 2, Rue & Held (2005)). Once again, we generated 30 sets of simulated data and fitted the models to each set in order to stabilize the variation due to generating the data.

Figure 4.4 displays interpolated surface plots of $\boldsymbol{y}_1 and \boldsymbol{y}_2$ for one set of simulated data. They provided evidence for the presence of spatio-temporal autocorrelation within the dependent variable. Spatio-temporal autocorrelation was exhibited for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2'$. For $\boldsymbol{y}_1$, there were clusters of higher values near the top right, bottom right, and center, and clusters of lower values displayed at the top, and bottom left. For $\boldsymbol{y}_2$, the spatial pattern of values changed from that of $\boldsymbol{y}_1$, indicating spatio-temporal autocorrelation. For $\boldsymbol{y}_2$, there was a larger cluster of low values at the bottom left compared to $\boldsymbol{y}_1$, and higher values to the right, and top of the plot. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}_t$ values for each t. Table 4.3 displays Moran's I and p-values for the two-sided test for the presence of spatial autocorrelation in $\boldsymbol{y}_t$ for each time point t for the first set of simulated data. We confirmed the presence of significant spatial autocorrelation within the dependent variable, for each time point, since the p-values were sufficiently small (less than 1%). Similar results were obtained for the four other sets of simulated data.

Table 4.3: Moran's I and p-values for the two-sided test for the presence of spatial autocorrelation.

Time, t	Moran's I, M_t	p-value
1	0.1106	6.33×10^{-10}
2	0.1085	1.02×10^{-9}

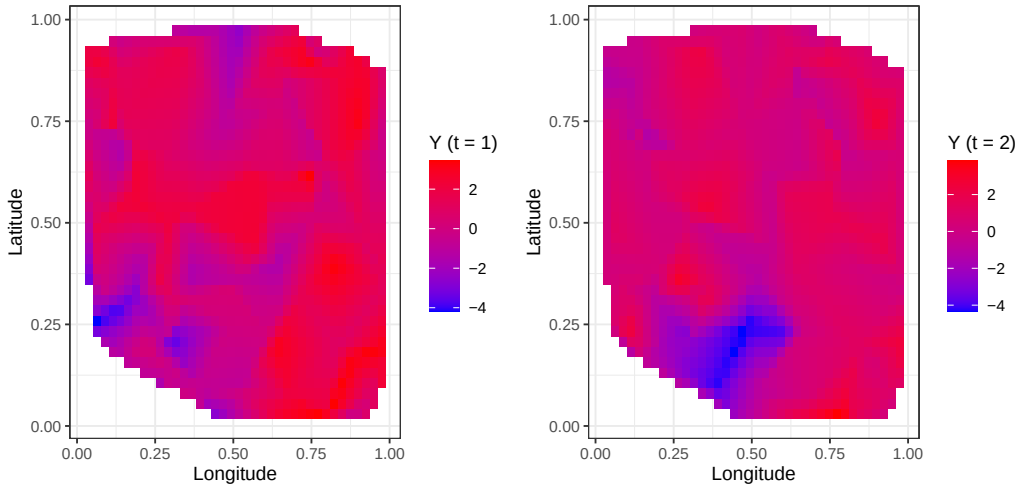


Figure 4.4: Interpolated surface plots of the simulated dependent variable. Spatio-temporal autocorrelation is exhibited for $(\boldsymbol{y}_1, \boldsymbol{y}_2)'$. For $\boldsymbol{y}_1$, there are clusters of higher values near the top right, bottom right, and center, and clusters of lower values displayed at the top, and bottom left. For $\boldsymbol{y}_2$, the pattern of values changes from that of $\boldsymbol{y}_1$, indicating spatio-temporal autocorrelation. For $\boldsymbol{y}_2$, there is a larger cluster of low values at the bottom left compared to $\boldsymbol{y}_1$, and higher values to the right, and top of the plot.

We fitted several CRN models to the simulated spatio-temporal data. Firstly, Model 1 is a CRN model that modelled the covariance matrix using the true network structure, obtained with a bandwidth d such that the network density is 5%. Models 2 – 10 were fitted next. They are a set of CRN models that assumed the network structure of the data was unknown and estimated by the adjacency matrix using Equation 4.8. Like the spatial simulation models, these models were determined by changing the band-

width, d, such that the approximating networks had densities of 1% – 4 % and 6% – 10% for Models 2 – 10 respectively. We estimated the covariance matrix for each model using Equation 4.27. Like the spatial simulation study, we also fitted a traditional exponential Matèrn covariance model, Model 11, in addition to fitting CRN models. For each spatio-temporal model, we produced posterior distributions of predicted values, $\boldsymbol{y}$, in order to calculate the measures of predicted accuracy and residual spatial autocorrelation. We also produced two further prosterior distributions of predicted values by Bayesian model averaging over Models 1 – 10 and over Models 2 – 10, following Section 4.5.1. We denoted these sets as BMA 1 and BMA 2. A description of each model and set of predicted values is given in Table 4.4. We describe each model in more detail in the following sections.

Model 1

Model 1, a CRN model (Lan et al., 2018; Liu et al., 2020), was fitted where the network structure of the data was assumed to be known and correctly specified. The network structure of the data was the same as the one used to generate the simulate data. The bandwidth d was such that the network had a density of 5%. The model is, described by Equations 4.25 – 4.27. The data likelihood is given as,

$$f(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\gamma}, \boldsymbol{A}) = \prod_{t=1}^{2} (2\pi)^{-\frac{200}{2}} |\boldsymbol{\Sigma}_t|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})' \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{y}_t - \beta_{0t}\mathbf{1}) \right\}, \tag{4.28}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)'$, $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})'$, $\boldsymbol{\gamma} = (\gamma_{01}, \gamma_{11}, \gamma_{21}, \gamma_{02}, \gamma_{12}, \gamma_{22})'$, and $\boldsymbol{\Sigma}_t = \gamma_{0t}\boldsymbol{I}_{200} + \gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2$, for $t = 1, 2$, and $\boldsymbol{A}$ is the same adjacency matrix used to simulate the data, defined by Equation 4.8, where $d = 0.2023$.

Table 4.4: Description of Models 1 – 11, BMA 1, and BMA 2 used to obtain the 13 sets of predictions of the spatio-temporal data.

Prediction set	Method	Network structure	Network density
1	Model 1	Correctly specified	5%
2	Model 2	Misspecified	1%
3	Model 3	Misspecified	2%
4	Model 4	Misspecified	3%
5	Model 5	Misspecified	4%
6	Model 6	Misspecified	6%
7	Model 7	Misspecified	7%
8	Model 8	Misspecified	8%
9	Model 9	Misspecified	9%
10	Model 10	Misspecified	10%
Prediction set	Method	Covariance structure	
11	Model 11	Exponential	

Prediction set	Method	Averaged over
12	BMA 1	Models 1 – 10
13	BMA 2	Models 2 – 10

Model 2 – 10

Models 2 – 10 were fitted next and are CRN models (Lan et al., 2018; Liu et al., 2020) where the network structure if the data was assumed to be unknown. We estimated the network structure using the adjacency matrix in the same way as Models 2 – 10 for the spatial simulation. The models are described by Equations 4.25 – 4.27 and the data likelihoods are given by,

$$f(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\gamma}, \boldsymbol{A}) = \prod_{t=1}^{2} (2\pi)^{-\frac{200}{2}} |\boldsymbol{\Sigma}_t|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})' \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{y}_t - \beta_{0t}\mathbf{1}) \right\}, \tag{4.29}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)'$, $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})'$, $\boldsymbol{\gamma} = (\gamma_{01}, \gamma_{11}, \gamma_{21}, \gamma_{02}, \gamma_{12}, \gamma_{22})'$, and $\boldsymbol{\Sigma}_t = \gamma_{0t}\boldsymbol{I}_{200} + \gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2$, for $t = 1, 2$.

Model 11

Model 11 was also fitted and is a traditional Matèrn covariance model with smoothness parameter $\nu = 0.5$, assuming temporal independence in the covariance matrix. The measurement equation for $\boldsymbol{y}_t$ under Model 11 is given by Equation 2.29, where the covariance matrix for each t is described as $\boldsymbol{C}_t = \boldsymbol{\Sigma}_t + \mathrm{T}$, where $\boldsymbol{\Sigma} = \frac{\sigma^2}{1-\rho^2}\boldsymbol{R}$ and $\mathrm{T} = \tau^2\boldsymbol{I}_{200}$. Model 11 assumed a temporally independent covariance function for the spatial process that depended on pairwise distances between locations. The function is defined by,

$$\boldsymbol{\Sigma}_t = \sigma^2 \exp(-\boldsymbol{D}/\psi), \tag{4.30}$$

where $\boldsymbol{D}$ is the same 200×200 matrix of pairwise distances between locations described in the spatial simulation, and is the same for $t = 1$ and $t = 2$. The parameter σ^2 is the spatial process variance and ψ is the spatial correlation strength parameter that measures the strength of correlation between two locations. The data likelihood is given by,

$$f(\boldsymbol{y}|\boldsymbol{\beta}_0, \sigma^2\tau^2, \psi, \boldsymbol{D}) = \prod_{t=1}^{2}(2\pi)^{-\frac{200}{2}}|\boldsymbol{C}_t|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\beta_{0t}\mathbf{1})'\boldsymbol{C}_t^{-1}(\boldsymbol{y}-\beta_{0t}\mathbf{1})\right\}. \tag{4.31}$$

BMA 1 and BMA 2

Like the spatial simulation, BMA 1 and BMA 2 represent the posterior distributions of predicted values that were calculated using Bayesian model averaging over Models 1 – 10 and over Models 2 – 10, respectively. For each set, we used the R package `bridgesampling`, which calculates the model evidence for each Model using Equation 4.9. We then compute the posterior model probabilities for Models 1 – 10 and Models 2 – 10, separately, using Equation 4.10. The posterior model probabilities were used

to calculate the weighted average posterior predictions, given by Equation 4.11.

We used MCMC to fit the models to each repetition of simulated data. For each model we assigned a vague prior to the intercept β_{0t}, for each t,

$$\beta_{0t} \sim \mathrm{N}(0, 1000).$$

For Models 1 – 10, we chose to assign the uniform priors to $\boldsymbol{\gamma}$ for the same reasons outlined in the spatial simulation. We assigned the following uniform priors, for each t, to minimise the impact of resulting posterior covariance matrices that would not be positive definite:

$$\begin{aligned}\gamma_{0t} &\sim \mathrm{U}(0, 5),\\ \gamma_{1t} &\sim \mathrm{U}(0, 2),\\ \gamma_{2t} &\sim \mathrm{U}(0, 1).\end{aligned}$$

The posterior distribution for Models 1 – 10 is,

$$f(\boldsymbol{\beta}_0, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{A}) \propto \prod_{t=1}^{2} |\boldsymbol{\Sigma}_t|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})\right\} \exp\left\{-\frac{1}{2}\frac{\beta_{0t}^2}{1000}\right\}, \tag{4.32}$$

where $\boldsymbol{\Sigma}_t = \gamma_{0t}\boldsymbol{I}_{200} + \gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2$.

For Model 11, we also assigned vague priors to the covariance parameters, the same as in the spatial simulation, σ^2, τ^2, and ψ,

$$\begin{aligned}\sigma^2 &\sim \mathrm{IG}(2, 1),\\ \tau^2 &\sim \mathrm{IG}(2, 1),\\ \psi &\sim \mathrm{IG}(2, 1).\end{aligned}$$

The posterior distribution for Model 11 is,

$$f(\boldsymbol{\beta}_0, \sigma^2, \tau^2, \psi|\boldsymbol{y}, \boldsymbol{D}) \propto \prod_{t=1}^{2} \left[|\boldsymbol{C}_t|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})'\boldsymbol{C}_t^{-1}(\boldsymbol{y}_t - \beta_{0t}\mathbf{1})\right\} \right.$$
$$\left. \exp\left\{ -\frac{1}{2}\frac{\beta_{0t}^2}{1000}\right\}\right] (\sigma^2)^{-3} \exp\left(-\frac{1}{\sigma^2}\right) (\tau^2)^{-3} \exp\left(-\frac{1}{\tau^2}\right) (\psi)^{-3} \exp\left(-\frac{1}{\psi}\right), \tag{4.33}$$

where $\boldsymbol{C}_t = \boldsymbol{\Sigma}_t + \mathrm{T} = \sigma^2 \exp(-\boldsymbol{D}/\psi) + \tau^2 \boldsymbol{I}_{200}$.

For the spatio-temporal simulation, each model was run for two chains of 7500 iterations. As with the spatial simulation study, we observed the chains converging to stationary distributions slowly and so 6750 (90%) of the iterations were discarded as warm-up. We thinned each chain by 2, to minimize autocorrelation in the posterior samples affording posterior draws of size 750. Trace plots, density curves, and autocorrelation plots were checked to determine that the posterior samples converged to stationary distributions. For conciseness, diagnostic plots were given for Model 1 only, in Figure B.2. In addition, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, given in Table B.2. Furthermore, the convergence diagnostics were reported for the first repetition of the simulation only. We noted similar convergence for each repetition. For each model fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary distributions, and appropriate exploration and mixing of the posterior distributions. Furthermore, $\hat{R}$ was found to be close to 1 for each parameter, indicating convergence (Brooks & Gelman, 1998).

Figure 4.5 and Table 4.5 display the posterior median RMSE (Equation 2.44), posterior median MAE (Equation 2.45), and posterior median Moran's I (Equation 2.8). In addition, the average estimated error in log scale for the covariance matrix, (COV, Equation 2.46) is displayed for each model, averaged over time. When we compared the CRN models, Models 1 – 10, we found that Models 2 – 10 had relatively similar posterior median RMSE and posterior median MAE. Further, the posterior median RMSE and MAE were smaller for Models 2 – 10 than for Model 1. We found this unusual, since Model 1 was the only model that used the correct network structure to generate the spatio-temporal data. This may be due to random chance in the simulation procedure.

When we compared the posterior median RMSE and MAE of Models 1 – 10 with that of the traditional Matèrn covariance model (Model 11), we

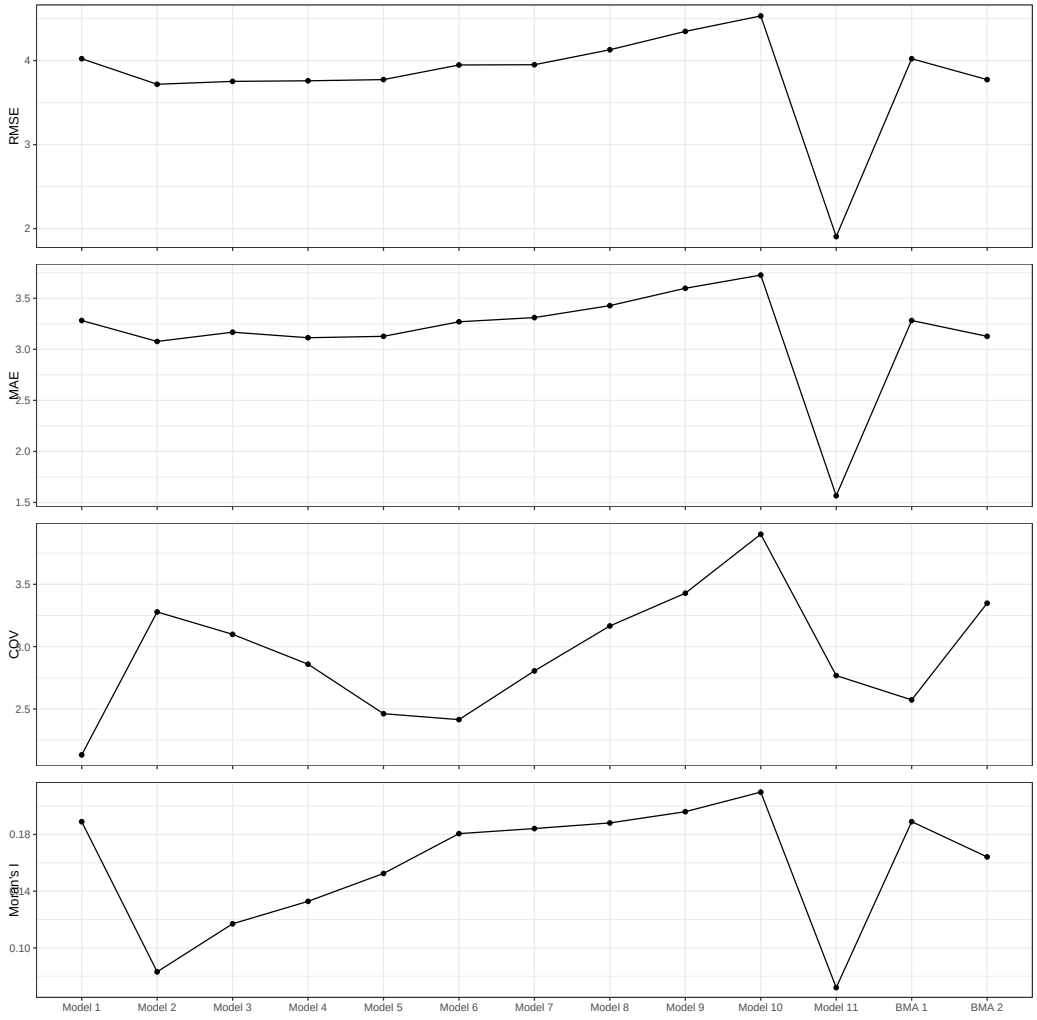


Figure 4.5: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I, summarised over the 30 sets of simulated spatio-temporal data.

Table 4.5: Medians of the posterior distributions for RMSE, MAE, COV, and Moran's I, summarised over the 30 sets of simulated spatio-temporal data.

Method	Predictive accuracy			Autocorrelation
for prediction	RMSE	MAE	COV	Moran's I
Model 1	4.022	3.282	2.132	0.189
Model 2	3.718	3.077	3.278	0.083
Model 3	3.752	3.168	3.099	0.117
Model 4	3.759	3.113	2.859	0.133
Model 5	3.774	3.127	2.462	0.153
Model 6	3.948	3.269	2.414	0.181
Model 7	3.950	3.311	2.805	0.184
Model 8	4.129	3.428	3.166	0.188
Model 9	4.348	3.598	3.428	0.196
Model 10	4.531	3.727	3.902	0.210
Model 11	1.905	1.566	2.768	0.072
BMA 1	4.022	3.282	2.573	0.189
BMA 2	3.774	3.127	3.348	0.164

observed some differences. For Models 2 – 10, the posterior median RMSE and MAE were larger than that of Model 11, suggesting that the CRN models were less accurate in terms of prediction. When we took the average posterior predicted values from Models 1 – 10 weighted by each models posterior model probability (BMA 1), we found the posterior median RMSE and MAE were much smaller than that from any other model. This suggests that averaging over a range of CRN models with different estimated network structures provides better predictive accuracy, compared to the individual CRN models and the Matèrn model. A similar result was seen when we averaged over Models 2 – 10 (BMA 2).

To assess each model's ability to account for spatial autocorrelation, we calculated Moran's I for each set of posterior residuals, using Equation 2.47. These were plotted in Figure 4.5, averaged over each time point t. Similar to the result of the spatial simulation study, we found that each

model was not able to account for spatial autocorrelation within the simulated data, with the exception of Models 2 – 4 and Model 11. For the majority of models, the posterior median of Moran's I was the same or larger than Moran's I for the simulated data, which was $I = 0.1106$ and $I = 0.1085$ for $t = 1$ and $t = 2$, respectively. Both BMA 1 and BMA 2 produced similar values for median Moran's I, and indicated that averaging over the models did not help account for more spatial autocorrelation. It was interesting to note that the value of Moran's I from fitting the Matèrn model was similar to the CRN models. This might indicate that the amount of spatial autocorrelation in the simulated data was too small for the models to account for.

We also measured accuracy of the models using COV. The posterior median COV is also displayed in Figure 4.5 for each model. Model 1 had, unsurprisingly, the lowest posterior median COV compared to all other models and sets of predicted values. We see a similar trend as in the spatial case, that as the estimated network structures moved closer in terms of network density to the true structure, the smaller the posterior median COV became. Model 11 had the highest posterior median COV of all models, highlighting that misspecification of the covariance matrix. BMA 1 and BMA 2 produced median posterior COV values similar to that of Model 1, and Models 5 – 7, respectively, showing that averaging can improve the accuracy of estimating the covariance matrix.

Moran's I was calculated for each set of posterior residuals for each model, and the absolute value for the posterior median Moran's I was plotted in Figure 4.5. We found that each model was able to account for some spatial autocorrelation within the simulated data. For each model, the absolute values of Moran's I were much smaller than Moran's I for the simulated data, which had values of 0.1106, and 0.1085 for $t = 1$ and $t = 2$, respectively. Across Models 1 – 10, we found a generally increasing trend in absolute posterior median Moran's I with increasing network density. This suggests that as the estimated network structures connectivity increased,

the amount of residual spatial autocorrelation increased. A similar trend was observed for the spatial simulation. This may be a reflection that these models were not able to capture the spatial autocorrelation because of model misspecification. Both Models 11 and 12, which were models averaged over Models 1 – 10 and Models 2 – 10 respectively produced Moran's I values that were similar to that of Models 6 – 10. This suggests that the individual CRN models were capable of accounting for spatial autocorrelation.

Another measure of accuracy is that given by the average estimated error (COV) between the model covariance matrix and the true covariance matrix. The posterior median COV is displayed in Figure 4.5 for each model. Of the six individual CRN models, Models 1 – 10, Model 1 had the lowest posterior median COV, which suggests that the posterior distribution of covariance matrices for Model 1 were the closest in Frobenius Norm to the true covariance matrix. As the network density of the estimated network structures increased, we observed a decreasing trend in the posterior median COV for Models 2 – 5, followed by a slight increasing trend for Models 6 – 10. This suggests that as the network structure approaches the true structure used to generate the data, the model covariance matrix becomes closer to the true covariance matrix. When we weighted the posterior distributions of the covariance matrix from Models 1 – 10, we observed a relatively low posterior median COV for Model 11 compared to Models 2 – 10. This was expected, since Model 1 contributed almost 40% to Model 11. The posterior median COV for Model 12 was higher than both that of Model 1 and Model 11, suggesting that the true network structure plays an important role in the accuracy of the model covariance matrix.

4.6 Case studies

4.6.1 New Zealand particulate matter

We performed a case study, in which we fitted several CRN models to the New Zealand particulate matter (PM10) concentration data described in Section 2.7.1. The aim was to find the best model to predict PM10 concentration across New Zealand, in terms of predictive accuracy RMSE, and MAE, while also accounting for spatial autocorrelation. Once a suitable model is chosen, we use it to produce an interpolated predictive map for particulate matter concentration, using covariate observations where particulate matter was not observed. Limited covariates were available to estimate the models, with only temperature (in °C) and wind speed (in m/s) considered. Temporal variation was not considered for this case study.

Mean PM10 recorded for the year 2013 were observed at 40 locations across New Zealand, denoted by $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_{40}))'$ (see Figure 2.2). Significant spatial autocorrelation was identified across the study region, as evidenced by the clusters of monitoring stations in the South Island of New Zealand that recorded higher concentrations of PM10 than stations in the North Island. This was confirmed by Moran's I, which was calculated as $I = 0.3577$ with a corresponding p-value for the two-sided test for presence of spatial autocorrelation of 3.23×10^{-8}.

We fitted ten CRN models on the PM10 data. Each model was determined by,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.34}$$

where $\boldsymbol{X}$ is a 40×3 design matrix of temperature and wind speed values observed at each location. The errors are modelled by a Gaussian process $\varepsilon \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Since the network structure of the monitoring stations that PM10 was observed was unknown, we estimated it using the adjacency matrix. For each model, we assumed that only connections between locations of path lengths 1 and 2 contributed to the covariance matrix. That is,

the covariance matrix was modelled by,

$$\boldsymbol{\Sigma} = \gamma_0 \boldsymbol{I}_{40} + \gamma_1 \boldsymbol{A} + \gamma_2 \boldsymbol{A}^2, \tag{4.35}$$

and we estimated the adjacency matrix, $\boldsymbol{A}$, using Equation 4.8 where the bandwidth, d, determined each of the ten models. The values of d were chosen such that each adjacency matrix resulted in a network structure with density from 1% – 10% using Equation 4.14.
For each model, we computed posterior distributions of predicted values, $\hat{\boldsymbol{y}}$, in order to calculate RMSE, MAE, Moran's I, and to choose a best model. We also produced a set of posterior predicted values by Bayesian model averaging over each of the ten models, using Equations 4.10 and 4.11.
Each Model assumed a data likelihood given by,

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{A}, \boldsymbol{X}) = (2\pi)^{-\frac{40}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}, \tag{4.36}$$

where $\boldsymbol{\Sigma}$ is given by Equation 4.35. We assigned a vague prior to the $\boldsymbol{\beta}$ parameters,

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, 1000 \boldsymbol{I}_{40}), \tag{4.37}$$

and vague uniform priors to $\boldsymbol{\gamma}$,

$$\gamma_0, \gamma_1, \gamma_2 \sim \mathrm{U}(0, 1000). \tag{4.38}$$

Therefore, the posterior distribution is,

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{A}, \boldsymbol{X}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \exp\left\{ -\frac{1}{2} \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{1000} \right\}. \tag{4.39}$$

We used MCMC to fit each model. Each model was run for two chains, each with 150000 iterations. To allow the chains to converge to stationary distributions we discarded 135000 iterations as warm-up. We thinned each chain by 3, to minimize autocorrelation in the posterior samples, affording posterior draws of size 30000. Trace plots, density plots, and autocorrelation plots were checked to determine that the posterior draws

converged to stationary distributions. For conciseness, we present the diagnostic plots for Model 1 only, in Figure B.3. In addition to the plots, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, given in Table B.3. For each model fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary distributions, and appropriate mixing of posterior distributions. Furthermore, $\hat{R}$ was calculated to be close to 1 for each parameter indicating convergence (Brooks & Gelman, 1998).

Figure 4.6 and Table 4.6 display the means of the posterior distributions for RMSE (Equation 2.44), MAE (Equation 2.45), and Moran's I (Equation 2.8). The model with the lowest posterior mean RMSE and MAE was Model 1. This model was based on a network structure with a relatively low number of connections between locations. When the number of connections in the network increased, we observed that the posterior mean RMSE increased and plateaued, with the exception of Model 9. This suggests that, generally, as the number of connections between PM10 monitoring stations network increased, the less accurate the models become. This might indicate that allowing observations separated by large distances to co-vary is inappropriate. When we averaged the posterior predicted values for Models 1 – 10, weighted by their posterior model probabilities, we observed mean RMSE and MAE that were similar to Model 9. This was because the set of posterior predicted values, BMA 1, were contributed to most by Model 9, which had a posterior model probability of 0.897 (Table 4.6).

Moran's I was calculated for each set of posterior residuals for each model, using Equation 2.47 and the posterior mean Moran's I was given in Table 4.6 and plotted in Figure 4.6. For each model fitted, we observed Moran's I lower than that of the raw data. This suggests that each model, based on estimating the network structure of the monitoring stations, was able to account for spatial autocorrelation. Spatial autocorrelation was accounted for most by Model 1, with Moran's I calculated as $I = 0.108$.

We chose Model 1 as the best model, based on posterior mean RMSE,

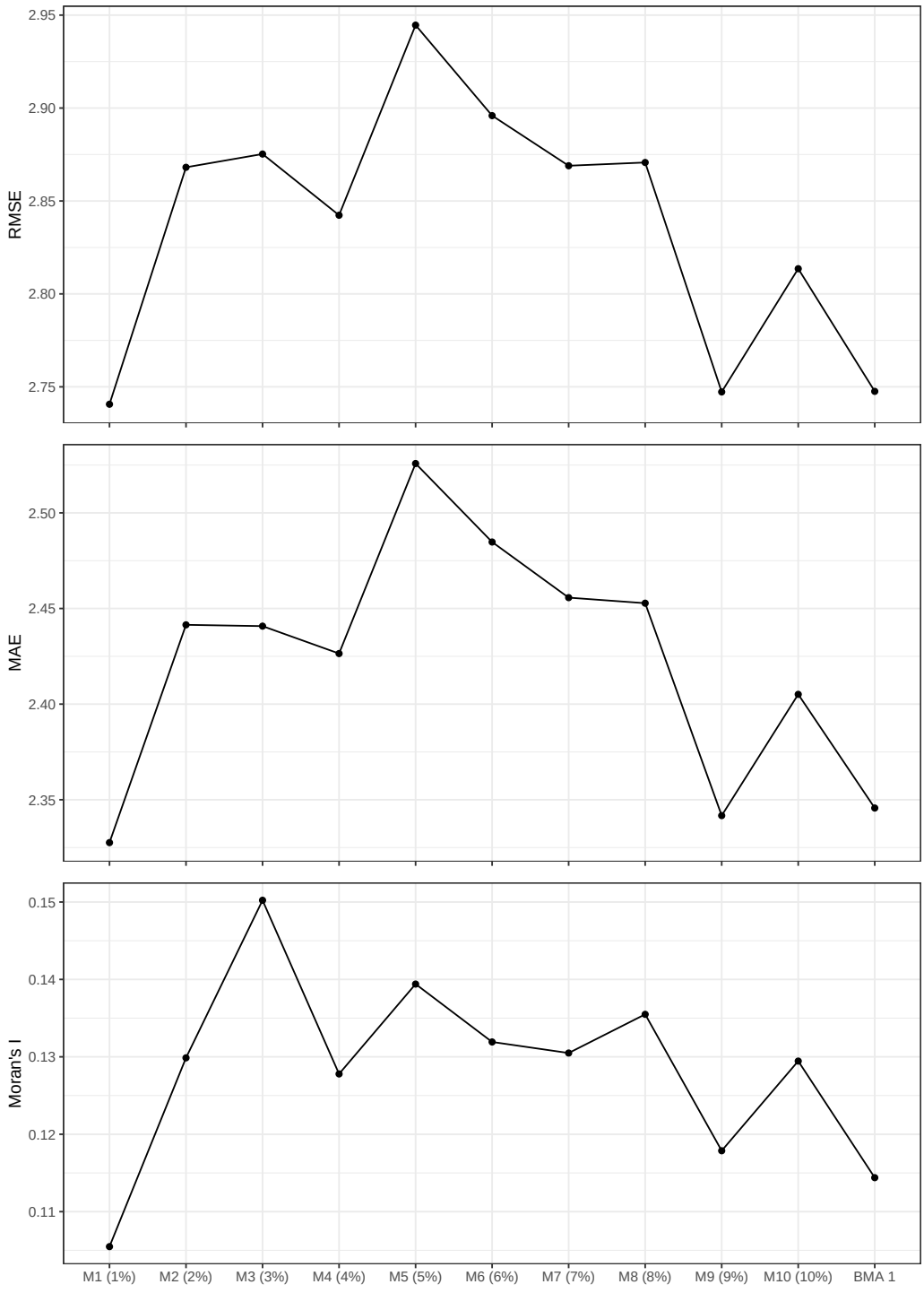


Figure 4.6: Means of the posterior distributions for RMSE, MAE, and Moran's I from models fitted to the PM10 concentration data.

Table 4.6: Means of the posterior distributions for RMSE, MAE, and Moran's I from models fitted to the PM10 concentration data.

Method	Predictive accuracy		Autocorrelation	Model probability
for prediction	RMSE	MAE	Moran's I	
Model 1	2.745	2.332	0.108	0.000
Model 2	2.862	2.432	0.135	0.000
Model 3	2.866	2.429	0.158	0.000
Model 4	2.822	2.406	0.125	0.000
Model 5	2.951	2.530	0.137	0.007
Model 6	2.882	2.471	0.124	0.004
Model 7	2.832	2.416	0.133	0.061
Model 8	2.861	2.448	0.130	0.031
Model 9	2.738	2.331	0.120	0.897
Model 10	2.809	2.403	0.126	0.000
BMA 1	2.737	2.334	0.116	

MAE, and Moran's I. In order to construct an interpolated surface map of PM10, we used temperature and wind speed observations from 347 locations across New Zealand in 2013 to calculate predicted values at new locations from Model 1. A map of the locations is shown in Figure 5.10. To compute the predicted values, we estimated the network structure of the new stations using the adjacency matrix and a bandwidth, d, such that the resulting network had a density of 1%. The predicted values were obtained through Bayesian kriging, given by Equation 2.43.

Figure 4.7 displays the interolated surface map of PM10 across New Zealand, using the CRN model with an estimated network structure with network density of 1%. We observed higher concentrations of PM10 in the northern part of the country, compared to lower values in the southern part. This finding does not agree with the observed PM10 concentrations used to fit the model, and suggests that CRN model may be overfitted. Furthermore, we doubt the appropriateness of estimating a different network structure for the new locations.

We now present the results from an application of the spatio-temporal

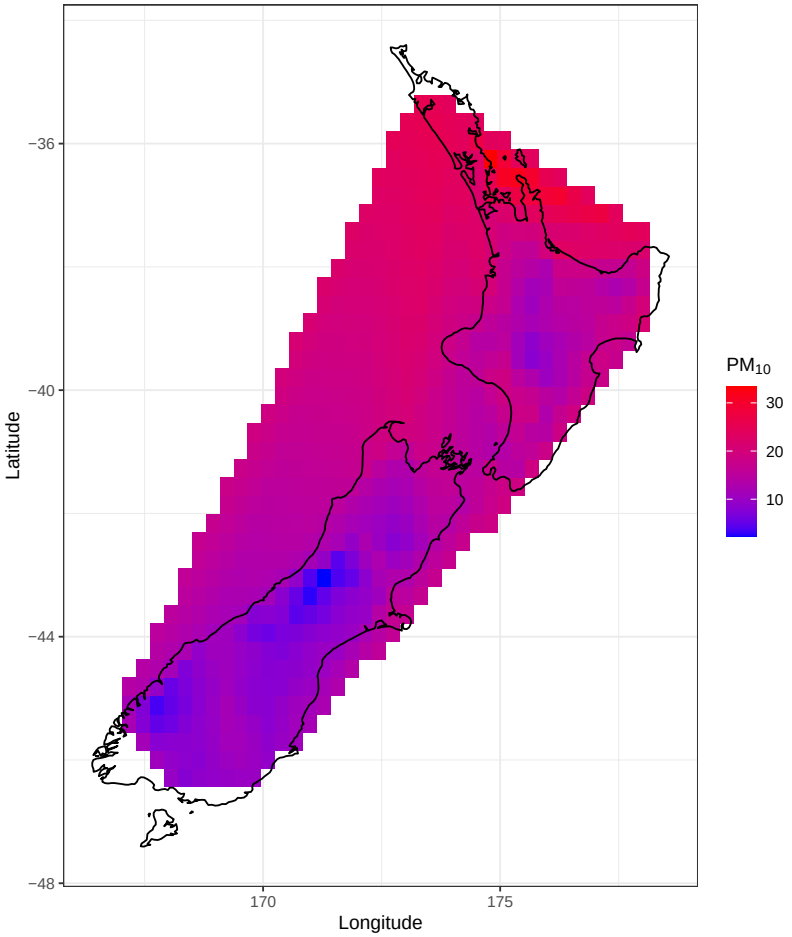


Figure 4.7: Interpolated surface plot of PM10 concentration generated from the posterior predicted values obtained by fitting Model 1 to the temperature and wind speed data.

CRN model to hoki catch weight data from the sub-Antarctic.

4.6.2 Sub-Antarctic hoki

In addition the New Zealand particulate matter case study, we performed a case study, in which we fitted several CRN models to the gridded sub-Antarctic hoki catch weight data described in Section 2.7.2. The aim was to find the best model to predict hoki catch weight across the sub-Antarctic region, in terms of predictive accuracy measures, RMSE and MAE, while also accounting for spatial autocorrelation.

Observed hoki catch weight in kilograms was recorded for 814 trawls taken throughout the sub-Antarctic region, for the years 2000 – 2008 (see Figure 1.4). The number of observations within each year changed, and the locations of the trawls were different each year. However, the CRN models that we developed in this thesis for point reference spatio-temporal data can only be applied to data that were observed at the same locations throughout time. Due to this fact, we gridded the hoki data according to the procedure detailed in Section 2.7.2, and fit the models to the mean hoki catch weight within the 38 grids, for years 2000 – 2008.

We fitted ten CRN models to the hoki data. Each model was determined by,

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \tag{4.40}$$

for $t = 1, ..., 9$ where $\boldsymbol{X}_t$ is a 38×2 design matrix of depth values observed at each grid center. The errors are modelled by a Gaussian process, $\boldsymbol{\varepsilon}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$. Since the network structure of the grid centers associated with observed hoki catch weight was unknown, we estimated it using the adjacency matrix. For each model, we assumed that only connections between locations of path lengths 1 and 2 contributed to the covariance matrix. Further, we assumed that the network structure did not change over time. The covariance matrix was modelled by,

$$\boldsymbol{\Sigma}_t = \gamma_{0t}\boldsymbol{I}_{38} + \gamma_{1t}\boldsymbol{A} + \gamma_{2t}\boldsymbol{A}^2, \tag{4.41}$$

and we estimated the adjacency matrix, $\boldsymbol{A}$, using Equation 4.8 where the bandwidth determined each of the ten models. The values of d were chosen such that the adjacency matrix resulted in a network structure with densities ranging from 1% to 10%, using Equation 4.14.
We computed the posterior distributions of the parameters using MCMC. Each model assumed a data likelihood given by,

$$f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{A},\boldsymbol{X}) = \prod_{t=1}^{9}(2\pi)^{-\frac{38}{2}}|\boldsymbol{\Sigma}_t|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})\right\}, \tag{4.42}$$

where $\boldsymbol{\Sigma}_t$ is given by Equation 4.41. We assigned a vague prior to the $\boldsymbol{\beta}$ parameters,

$$\beta_{0t},\beta_{1t} \sim \mathrm{N}(\boldsymbol{0}, 1000\boldsymbol{I}_{38}), \tag{4.43}$$

and vague uniform priors to $\boldsymbol{\gamma}$,

$$\gamma_{0t},\gamma_{1t},\gamma_{2t} \sim \mathrm{U}(0,1000). \tag{4.44}$$

Therefore, the posterior distribution is,

$$f(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\boldsymbol{A},\boldsymbol{X}) \propto \prod_{t=1}^{9}|\boldsymbol{\Sigma}_t|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})\right\}\exp\left\{-\frac{1}{2}\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{1000}\right\}. \tag{4.45}$$

Each model was run for two chains, each with 75000 iterations. To allow the chains to converge to stationary distributions, we discarded 67500 iterations as warm-up. We thinned each chain by 2, to minimize autocorrelation in the posterior samples, affording posterior draws of size 7500. Trace plots, density plots, and autocorrelation plots were checked to determine that the posterior draws converged to stationary distributions. For conciseness, we present the diagnostic plots for Model 1 only, in Figures B.4 – B.7. In addition to the plots, we calculated the potential scale reduction factor, $\hat{R}$, for each parameter, given in Tables B.4 – B.7. For each model fitted to each set of data, the diagnostic plots showed sufficient evidence of convergence to stationary distributions, and appropriate mixing of posterior distributions.

For each model, we computed the posterior distributions of predicted values, $\hat{\boldsymbol{y}}$, using Bayesian kriging (Equation 2.43). For Models 4 and 7, we were not able to compute the posterior distributions of predicted values, because the posterior distributions for the γ parameters afforded covariance matrices that were not positive definite. As a result, we excluded these models from comparison.

We used the predicted values to calculate RMSE, MAE, and Moran's I. We also produced a set of posterior predicted values by Bayesian model averaging over each of the eight models with calculated predicted values, using Equations 4.10 and 4.11. Figure 4.8 and Table 4.7 display the means of the posterior distributions for RMSE, MAE, and Moran's I on the residuals, averaged over time. There is a clear trend in the mean RMSE and MAE with increasing connectivity of the estimated network structure. As the bandwidth increased to allow more connections in the estimated network structure, the higher the mean RMSE and MAE for each model. This suggests that allowing observations separated by larger distances to covary does not improve predictive accuracy. Averaging over all models weighted by posterior model probability, we improve the predictive accuracy, seen as a decrease in mean RMSE and MAE.

Model 1 was able to account for the most spatial autocorrelation, reflected in having mean Moran's I closest to 0. As connectivity increased, so did the mean Moran's I. This suggests that allowing observations separated by larger distances to co-vary does not necessarily account for spatial autocorrelation. Further, averaging over all models did not improve Moran's I on the residuals.

Model 1 appeared to perform the best in terms of predictive accuracy and accounting for spatial autocorrelation. This suggests that the covariance matrix was not as important in improving accuracy and accounting for spatial autocorrelation as the temporally dependent mean function was.

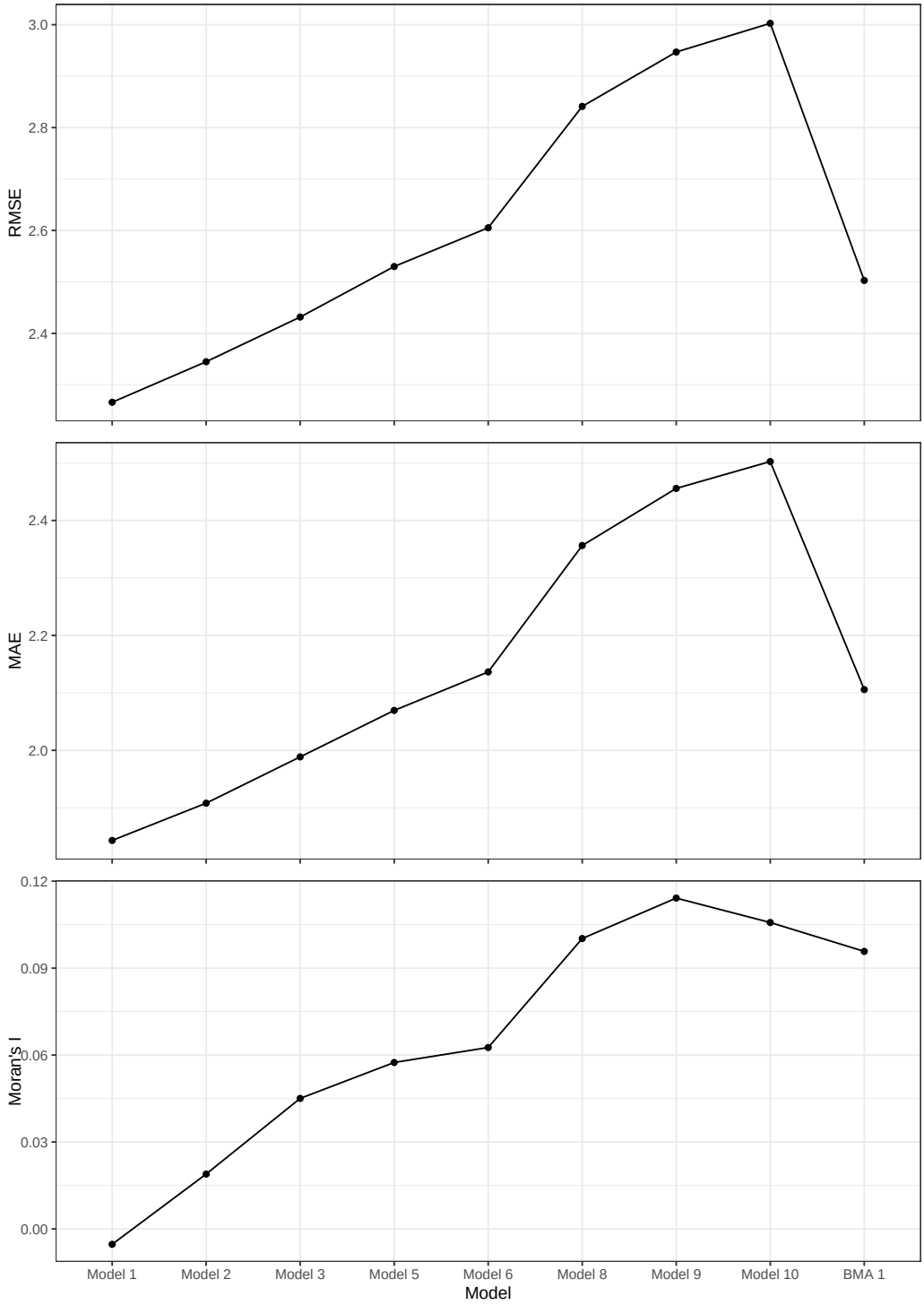


Figure 4.8: Means of the posterior distributions for RMSE, MAE, and Moran's I from models fitted to the hoki catch weight data.

Table 4.7: Means of the posterior distributions for RMSE, MAE, and Moran's I from models fitted to the hoki catch weight data.

Method	Predictive accuracy		Autocorrelation	Model probability
for prediction	RMSE	MAE	Moran's I	
Model 1	2.266	1.843	-0.005	0.001
Model 2	2.345	1.908	0.019	0.000
Model 3	2.432	1.989	0.045	0.285
Model 5	2.530	2.069	0.057	0.000
Model 6	2.605	2.136	0.063	0.000
Model 8	2.841	2.356	0.100	0.000
Model 9	2.947	2.456	0.114	0.714
Model 10	3.003	2.503	0.106	0.000
BMA 1	2.503	2.106	0.096	

4.7 Conclusion

Covariance regression network (CRN) models were proposed for spatial and spatio-temporal point referenced data in this chapter. They were shown to provide more flexibility in modelling the covariance function of spatial and spatio-temporal processes. The best results in terms of predictive accuracy measures, RMSE and MAE, were obtained when we performed Bayesian model averaging over the CRN models that were based on different estimates of the network structure.

Chapter 5

Geographic random forest for spatial and spatio-temporal data

The 21st century has seen an increasing interest in the use of computationally intensive and primarily data driven algorithms. These techniques, known collectively as machine learning, have seen growth in many directions with wide ranging application in data mining, pattern recognition, regression, and classification problems (Hengl 2018). Machine learning has always been concerned with the understanding and uncovering of complex relationships in data. Not only is there a need to produce accurate predictions, but also the ability to recover knowledge in an intelligible way.

In the field of spatial and spatio-temporal statistics, spatial prediction is a key goal. That is, the prediction of the occurrence, value, and/or state of geographically varying phenomena. As seen in previous chapters of this thesis, common methods for spatial prediction involve model-based approaches, such as linear regression, kriging, or a combination of the two. However, model-based methods make strong assumptions about the data generation model and sampling method, which may not always be appropriate. Some of the assumptions that are most frequently made are that: the residuals are normally distributed; the residuals are stationary;

the response variable is linearly related to predictors; and, that the model is correct. Furthermore, model-based approaches can be computationally intensive, usually requiring inversion of a covariance matrix. This issue worsens as the number of locations and time points increases.

Machine learning algorithms (MLAs) are non-parametric. They make no assumption about an underlying generative model. Further, there is less focus on inferring data generation processes. Instead, the focus is on developing a procedure for accurate predictions. MLAs have a benefit over parametric models in that they are able to handle high dimensional and highly correlated data (Schratz et al., 2019), as well as make no assumptions about sampling (Hengl et al., 2018). However, MLA's are considered a "black box" methodology, in that there is reduced interpretability. Despite this, there is increasing interest in using MLAs in the field of spatial and spatio-temporal statistics.

Throughout the machine learning literature, tree-based methods stand as one of the most effective and useful techniques that can produce both reliable and interpretable results, on mostly any kind of data (Loupe, 2014). The success of tree-based methods is defined by several properties. They are non-parametric and can model arbitrarily complex relationships between inputs and outputs, without any *a priori* assumption; handle heterogeneous data (ordered or categorical variables, or a mixture of both); can intrinsically implement feature selection, making them robust to irrelevant or noisy variables to some extent; are robust to outliers or errors in labels; and, they are easily interpretable.

The most popular tree-based MLA is random forest (RF) proposed by Breiman (2001). The popularity of RF is reflected by the 55232 citations of Breiman's paper as of February 2020. The application of RF to spatial, and spatio-temporal data is becoming more frequent (Georganos et al., 2019). However, RF is not a spatial technique, and as a result, does not explicitly take into account any spatial autocorrelation within the variable of interest.

There have been many attempts to incorporate spatial autocorrelation into MLAs, for the purpose of spatial prediction. In this chapter, we explore a variety of MLAs and their application to spatial prediction. We review RF approaches in particular, and see how the traditional RF methodology has been modified to combat the issue of unaccounted spatial autocorrelation. Section 5.1 provides a literature review that details the successes and shortcomings of several MLAs applied to spatial data. In Section 5.2 we take a deeper look at RF and provide examples from the literature of its application to spatial and spatio-temporal data. In Section 5.3, we introduce the geographic RF methodology proposed by Georganos et al. (2019). Two subsections are dedicated to the development of a cluster based technique and to the extension of the geographic RF to spatio-temporal data. In Section 5.4, we propose a geographic random forest methodology that uses a neighbourhood structure for each location based on a clustering algorithm. Clustering based on geographic measures such as longitude and latitude would ensure that observations that may be too far away to have any influence on the observations near the locations where a local RF is fit are not selected to form the neighbourhood. In Section 5.5, we propose an extension of geographic random forest to model spatio-temporal data in order to allow prediction of a dependent variable that has been observed at fixed locations at fixed time points. Section 5.6 outlines a simulation study that compares the geographic RF methodology for different settings. The methods are applied to two datasets in Section 5.7 and concluding remarks are made in Section 5.8.

5.1 Literature review

The term **data mining** is often associated with MLA. Data mining is the practice of examining large pre-existing databases in order to generate new information, whereas MLAs are automatic and learn from the data, in addition to embodying the same principles as data mining. Within the

data mining field, it has been stressed that the presence of spatial autocorrelation within data requires an appropriate treatment to deal with its effects. LeSage & Pace (2001) have shown that the inclusion of spatial autocorrelation of the dependent variable in a data mining application provided an improvement in fit. A reason for this might be that spatially autocorrelated data violate common assumptions of model-based methods, such as independence (Legendre, 1993). Furthermore, it allows us to capture complex phenomena within data, such as non-stationarity.

Several studies have examined the effect of spatial autocorrelation in a data mining setting. These studies involved the use of MLAs, and attempted to account for the inherent spatial autocorrelation for a range of uses, such as spatial clustering, classification, regression, and relational data mining.

A foundation for incorporating spatial autocorrelation into data mining methodology was laid by Huang et al. (2004). They proposed and empirically validated a data mining method for predicting the colocation patterns of geographic objects at particular locations. The method was based on logistic regression and Bayesian classification that explicitly takes spatial dimension into account.

In the realm of spatial clustering and classification, Scrucca et al. (2005) proposed a clustering procedure for identifying spatial clusters based on the contiguity structure of objects and their attribute information. This was implemented using K-means clustering to incorporate spatial structure through measures of spatial autocorrelation, such as Moran's I and Geary's C.

In predictive data mining, Li & Claramunt (2006) used "Spatial entropy" to capture autocorrelation in order to adapt classification trees for handling geographical data. Also in predictive data mining, Bel et al. (2009) modified Breiman's classification trees to take into account the irregularity of sampling by weighting the data according to their spatial pattern. This was carried out by using Voronoi tessellation.

In the spatio-temporal classification literature, Zhang et al. (2003) used a spatial autocorrelation-based search tree to solve the problems of correlation-based similarity range queries that are used to identify pairs of potentially interacting elements from the cross product of two spatial time series datasets. The algorithm divided a collection of time series into hierarchies based on spatial autocorrelation to facilitate similarity queries and joins. Further, Zhang et al. (2003) proposed processing strategies for correlation-based similarity range queries and similarity joins using the proposed spatial autocorrelation-based search trees.

Stojanova et al. (2011) proposed an MLA that explicitly considered spatial autocorrelation when building the algorithm. The method was based on predictive clustering trees (PCTs) and was able to combine the possibility of capturing global and local effects at different levels of the tree. PCTs combine elements from both prediction and clustering. It is a form of supervised learning, with a predictive algorithm assigned to each cluster. The main assumption made was that if there was high spatial autocorrelation between observations in the dataset, then not only would the observations have similar target values but they would also likely be in the same spatial neighbourhood. PCTs were shown to offer a unique opportunity to increase the accuracy of the predictive algorithms without performing spatial partitioning that could lead to losing generality of the induced models.

Random forest (RF, Breiman (2001)) is an MLA also based on decision trees. It has been demonstrated to be a promising technique for spatial prediction within data mining (Nussbaum et al., 2018; Prasad et al., 2006; Hengl et al., 2015, 2018). However, the spatial locations of the observations are ignored by the algorithm and hence, spatial autocorrelation is not taken into account.

The modelling of covariates and spatial autocorrelation jointly using MLAs is a relatively sparse area of research (Hengl et al., 2018). Hengl et al. (2015) compared random forest and linear regression models on soil fertility in

Africa. A 5-fold cross validation demonstrated that the random forest algorithm consistently outperformed the linear regression algorithm by producing more accurate predictions resulting in lower RMSE. However, it was shown that the RF was sensitive to artifacts in the input data. Furthermore, the algorithms did not account for spatial autocorrelation, leading to spatially autocorrelated residuals.

Requia et al. (2019) compared three approaches for fitting models to PM2.5 concentrations from Eastern Massachusetts in the United States. The approaches tried were ordinary krigging, a hybrid methodology (geographical interpolation and land use regression), and RF. While it was found that the hybrid method that combined kriging with land use regression performed better than the ordinary kriging model in terms of capturing spatial variation (by accounting for spatial autocorrelation), the RF algorithm found a substantial improvement in terms of R^2, and RMSE. Requia et al. (2019) attributed this to the kriging and regression models' limited capacity to account for the complex relationship between variables since independence of observations, the predictor distributions, and collinearity can significantly impact these methods, while the RF approach finds the importance of those features trivial. The increase in performance between the hybrid method and the random forest method is shown as a considerable increase in the explained concentration variance for all PM2.5 components.

Prasad et al. (2006) compared four MLAs in the context of tree species distribution modelling under future climate impacts. They sought to compare RF to regression tree analysis Lewis (2000), bootstrap aggregated (bagged) trees, and multivariate adaptive regression splines (MARS). In essence, bagged trees is an ensemble method built up from multiple regression trees, while RF is an adaptation of bagged trees. The MARS method is different, in that it is not built from regression trees. It was found that bagged trees and RF had a distinct advantage over MARS and regression tree analysis in predictive mapping. They were found to be

more effective than single regression tree outputs, because they produced more accurate predictions. RF was demonstrated to be superior for this type of application because it provided a smoother response surface in that the tree importance values (IVs) graded smoothly from lower to higher values and there was no jumping of classes. Furthermore, Prasad et al. (2006) did consider not accounting for spatial autocorrelation.

In Hengl et al. (2018), RF was fit to daily precipitation measurements from Boulder, Colorado, a spatio-temporal data set. Distance in the time domain was represented by cumulative days since 1970 and day of the year, to capture long term trends and seasonality effects respectively. In addition to the two time variables included as features in the RF, elevation maps, and long term precipitation maps, were included as geographic measures. It was found that the most important variables for predicting daily precipitation were both time covariates. Hengl et al. (2018) compared their RF output to that from a traditional geostatistical kriging model, applied to the same data. Both approaches gave comparable results in terms of prediction accuracy, however, it was stated that the RF method was able to reflect more closely influence of relief and impact of individual stations on predictions, and map prediction errors with higher contrast.

We wish to investigate RF approaches in greater depth. In particular, approaches that also explicitly account for spatial autocorrelation. Furthermore, we identify a gap in the MLA literature in the form of a lack of RF methods for spatio-temporal data that account for spatial and temporal autocorrelation. In the next section, we define the RF algorithm, first proposed by Breiman (2001). We review the current state of the literature surrounding RF and pay particular attention to those that attempt to incorporate spatial autocorrelation.

5.2 Random forest

Random forest (RF, Breiman (2001); Prasad et al. (2006); Biau & Scornet (2016)) is a supervised MLA that is an extension of bootstrap aggregated (bagged) trees. It was developed in order to improve on the over fitting that resulted from using a single classification and regression tree (CART) to make predictions. Given a training data set, a random subset is sampled with replacement, and is used to construct a decision tree based on a subset of explanatory variables. The remaining portion of training cases are put through the tree and are classified. Any cases that were misclassified are then used to grow the tree further, and the process is repeated. Once all cases are accurately classified, the tree is complete.

Of the training set, approximately one third is kept out of the construction of the tree, and denoted as the out of bag (OOB) data. At each decision node within each tree, a random selection of features are selected. This was shown to improve on the issue of bias (Breiman, 2001). The OOB data is used to get a classification error rate as trees are added to the forest and to measure input variable (feature) importance. In the end, a sample can be classified or predicted using the majority vote (classification) or the average prediction (regression), respectively, over all trees in the forest, similar to the bagging concept.

A variety of studies have demonstrated that it is one of the best MLAs currently available (Cutler et al., 2007; Boulesteix et al., 2012; Fox et al., 2017). However, RF is considered a non-spatial approach to spatial prediction. The locations that were sampled and the general sampling pattern are ignored during the estimation of the MLA parameters. This can potentially lead to less than optimal predictions and systematic under- and over-prediction. These problems are further purported when spatial autocorrelation within the response variable is high (Hengl et al., 2018).

Hengl et al. (2018) provided a possible solution to the problem of applying RF to data of a spatial nature. They suggested that the solution lies

within preparing geographical measures of proximity and connectivity between observations. A list of these geographical measures were provided and include: geographical coordinates such as longitude and latitude; Euclidean distance to reference points, such as distance to the center or edge of the study region; Euclidean distance to sampling locations, such as distances from observation locations or other distance measures; downslope distances; resistance distances, such as distances of the cumulative effort derived using terrain ruggedness and/or natural obstacles. We refer to the approach of incorporating geographical measures as features in RF as spatial RF.

Hengl et al. (2018) compared the performance of a "state-of-the-art" geostatistical model-based approach to their spatial RF approach, which included geographical measures, implemented through the ranger R package. The comparison was carried out on the Meuse dataset, available through the sp R package. Particular focus was placed on mapping zinc (Zn) concentration. An assumption was made that concentration of metals in soil (such as Zn) is controlled by river flooding as sediments are carried upstream. Geographical buffer distance was included as the only geographical covariate. The overall pattern of the spatial map by model-based and spatial RF approach were similar. Smoothing was more prominent in the case of the spatial RF approach, and was concluded to be a result of the averaging of trees in the random forest. The overall correlation between the model-based and spatial RF maps was high (r=0.97). Cross validation was used to assess model performance and it showed that the model-based approach was more accurate in terms of R^2 than the spatial RF approach. There was no remaining spatial autocorrelation in the residuals in both cases, and it was concluded that both methods had fully accounted for the spatial structure in the data.

When more geographical covariates were included, a reduction in MSE was observed. In addition, a small difference in spatial patterns were observed between this model and the model with one geographical covari-

ate. Hengl et al. (2018) concluded that geographic buffer distance was the covariate most important for mapping Zn concentration.
MLAs such as random forest have been shown to increase the accuracy of predictions when compared to other MLAs and model-based methods. In addition, their non-parametric nature means no assumption on the nature of sampling or the structure of the data need be imposed. The spatial RF methodology of Hengl et al. (2018) that involved incorporating geographic measures as features displayed promising results. Predictions from the spatial RF were accurate predictions and analysis of the residuals showed that autocorrelation had been taken into account. However, there may be reservation. Including geographic measures as explanatory variables in RF algorithm may not necessarily capture complex spatial phenomena such as non-stationarity. We wish to explore another avenue for accounting for spatial autocorrelation when using an RF approach, that is also able to capture variations in spatial autocorrelation throughout the study area. Georganos et al. (2019) presented a novel geographical implementation of RF, so-called geographical random forest (geoRF) for both prediction and exploration to model population as a function of remote sensing covariates.

5.3 Geographical random forest

Georganos et al. (2019) presented a novel geographical implementation of RF, so-called geographical random forest (geoRF) for the purposes of both prediction, and exploration. The objective of their study was to model population as a function of remote sensing covariates. The methodology can be described as a disaggregation of RF into geographical space in the form of local sub models. The motivation behind this was loosely based on the model-based concept of spatial varying coefficient models, otherwise known as geographic weighted regression (GWR,Fotheringham et al. (2003)). In essence, for each location i, a local RF is computed but only in-

cluding n number of nearby observations. This would lead to the calculation of an RF for each training data point, each with its own performance, predictive power, and feature importance (Georganos et al., 2019).

The area that the local sub model operates within is called the neighbourhood, or kernel. The maximum distance between a data point and its kernel is called the bandwidth (Brunsdon et al., 1998). Georganos et al. (2019) considered one of two main types of kernel, 'adaptive', and 'fixed' (Kalogirou, 2016). An adaptive bandwidth defines a kernel by the n nearest neighbours, while a fixed bandwidth defines a kernel by a circle whose radius is the bandwidth (Brunsdon et al., 1998; Fotheringham et al., 2003).

Figure 5.1 presents an illustration of the observations that are selected to be part of the neighbourhood of two locations (the crosses in bold) when a fixed bandwidth is used, and when an adaptive bandwidth is used. When an adaptive bandwidth of $n = 2$ is used, we see that the two nearest observations to each of the locations in bold are selected to form the neighbourhood of their respective local RF. For each location, three observations would be used. The neighbourhood of the leftmost location in bold displays more sparsity than that of the rightmost location in bold. In otherwords, the observations are closer on average in the rightmost location in bolds neighbourhood than those in the leftmost location in bold. When the bandwidth is fixed at distance d, we see that observations within the green circles of radius d around the locations in bold are selected to form the neighbourhood of their respective local RF. Because the density of locations changes across the study region (which is likely in reality), there are different numbers of observations selected to form the neighbourhoods for each location. In general, the more dense the observations are around a location, the more observations are selected to form the neighbourhood and vice versa. Georganos et al. (2019) chose to use the adaptive kernel because of its apparent advantage when sampling density differs across space. However, a disadvantage might be that some observations may be selected to form the neighbourhood when they are too far away to have

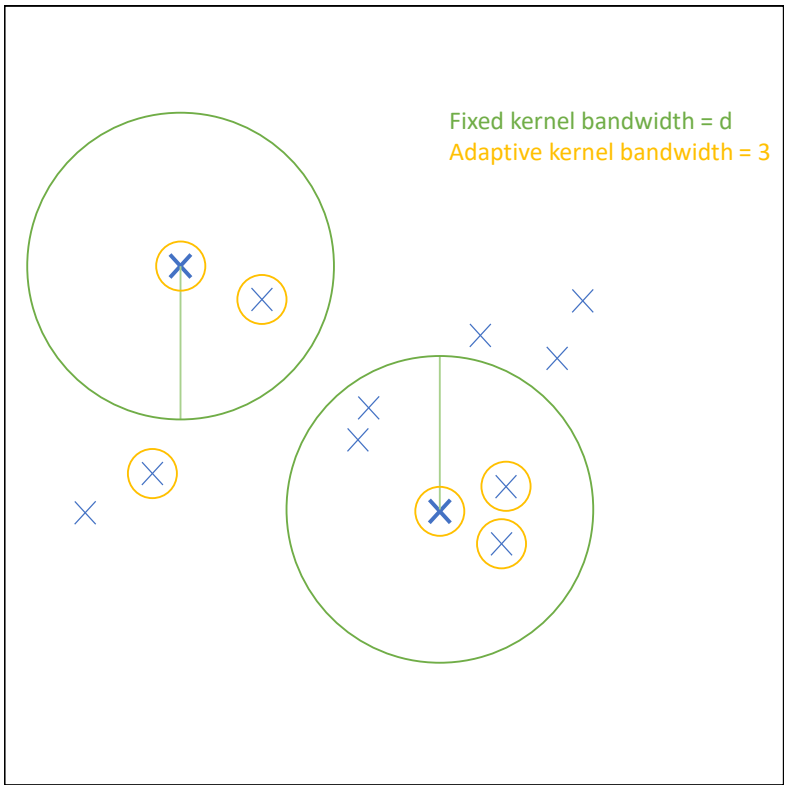


Figure 5.1: Illustration showing the neighbourhoods around two locations using an adaptive bandwidth, and a fixed bandwidth to define the neighbourhood. An adaptive bandwidth selects the nearest n (in this example, 3) observations to each location, shown in yellow circles. A fixed bandwidth selects all the observations within a circle (green) of radius d. Each method results in a different number of observations being selected to build the local random forests.

any influence on the observations near the location where the RF is built. We attempt to address this issue later in the next section.

In order to generate predicted values, the predictions from the global forest and local forests are combined using a weight parameter, a, according to,

$$\hat{\boldsymbol{y}} = a\hat{\boldsymbol{y}}_l + (1-a)\hat{\boldsymbol{y}}_g \tag{5.1}$$

where $\hat{\boldsymbol{y}}$ is the vector of predicted values, $\hat{\boldsymbol{y}}_g$ is the vector of predicted values from the global forest, and $\hat{\boldsymbol{y}}_l$ is the vector of predicted values from the local forests. Georganos et al. (2019) claim that fusing the predictions allows for an extraction of the locally heterogenous signal from the local sub model, which contributes low bias, and merges it to that of a global model that uses more data, and hence contributes a low variance. The weight parameter, a, is user defined between 0 and 1. A weight parameter of $a = 0$ corresponds to zero contribution from the local sub models, and only the global model being used to compute the predictions. This is essentially no different from fitting an RF to the entire data. A weight parameter of $a = 1$ corresponds to only contribution from the local sub models. For predicting on new spatial locations, the closest available local sub model was used.

In order to evaluate the algorithm, the root mean squared error (RMSE), mean absolute error (MAE), and Moran's I were calculated. The first two quantities measured the accuracy of prediction of the algorithm compared to the test data set, while Moran's I was used to ascertain whether autocorrelation still exists within the residuals. Georganos et al. (2019) compared MAE, RMSE and Moran's I for a variety of geoRFs applied to a population census dataset at the neighbourhood scale in Dakar, Senegal using land cover (LC) classification products to train the models. Here, LC is the observed physical cover on the Earth's surface. Four geoRF designs were constructed. The first included all LC classes and geographical coordinates as explanatory factors and was referred to as LC_XY. The second included all LC classes as explanatory factors, referred to as LC_. The third included three types of built-up LC classes and geographical coordinates as explanatory factors, 3BU_XY. Finally, the fourth included three types of built-up LC classes as input, 3BU_. Each geoRF design was defined by an adaptive kernel, constructed using different bandwidths ranging from 100 to 1100. Three weight parameters were also chosen to be compared, $a = 0.25, a = 0.5$, and, $a = 0.75$. A traditional RF was also fit for compar-

ison. A pattern was identified in the distribution of RMSE and MAE as a function of the bandwidth and weight parameter specified, irrespective of the geoRF model used. It was found in all four modelling designs that weighting the local models too heavily (when a = 0.75) was not optimal in terms of accuracy. In most cases, it was stated that a global RF model would perform similarly or better. However, when the weighting toward the local models were decreased (a = 0.5, and a = 0.25), the GRF was found to produce better predictions for some cases (when the bandwidth ranged between 100 and 400). The method found to work the best according to RMSE and MAE was that of $3BU_XY$ geoRF, with weight parameter of 0.25, and a bandwidth of 400, with its global counterpart underperforming.
We wish to investigate the impact of fixed and adaptive kernels in addition to bandwidth and local weighting effects on the measures of model accuracy (RMSE, MAE). Furthermore, we wish to investigate how much residual spatial autocorrelation is left over after fitting the random forest.

5.4 Cluster approach

The choice of neighbourhood structure (adaptive vs. fixed) and bandwidth could be considered arbitrary. To find the optimal bandwidth, we have seen in Georganos et al. (2019) that multiple geoRFs need to be constructed on the data. A neighbourhood structure could potentially be built more objectively by using a clustering algorithm. Clustering observations based on location or other spatially correlated covariates could allow for smaller groups of observations that have similar spatial autocorrelation structure. We believe that using clusters would be advantageous when non-stationarity is present within the data. Figure 5.2 gives an illustration of the neighbourhoods for two locations selected using clustering.
We propose a geographic random forest methodology that uses a neighbourhood structure for each location based on a clustering algorithm. A training data set is clustered based on geographic measures, such as longi-

tude and latitude. This would ensure that observations that may be too far away to have any influence on the observations near the locations where a local RF is fit are not selected to form the neighbourhood. An RF is then fit to each location using all the observations belonging to the cluster that the location belongs to. A global RF is constructed as is the case in Georganos et al. (2019). Identically, to obtain the predicted values of a response variable for new locations, the covariates are run through the local RF that correspond to the locations closest to them. The covariates are also run through the global RF and the final predictions are computed as a weighted average of the local forests and global forest.
We propose the use of the K-means clustering algorithm Steinhaus (1956); MacQueen et al. (1967) to cluster the observations based on a geographic distance measure. The number of clusters was selected, using the "elbow" method Marutho et al. (2018). Clustering in this way, the neighbourhoods for each local random forest might include observations from locations different from those selected using an adaptive or fixed bandwidth. As such, we believe clustering would provide a more flexible and objective way of selecting neighbourhoods for the local random forests. Further, if clustered using an appropriate distance measure, more spatial features may potentially be incorporated into the local RF at each location. This would not only account for spatial autocorrelation, but would also take into account the existence of non-stationarity.

5.5 Geographical random forest for spatio-temporal data

Geographic random forest has not yet been extended to model spatio-temporal data. We present this extension to allow prediction and explo-

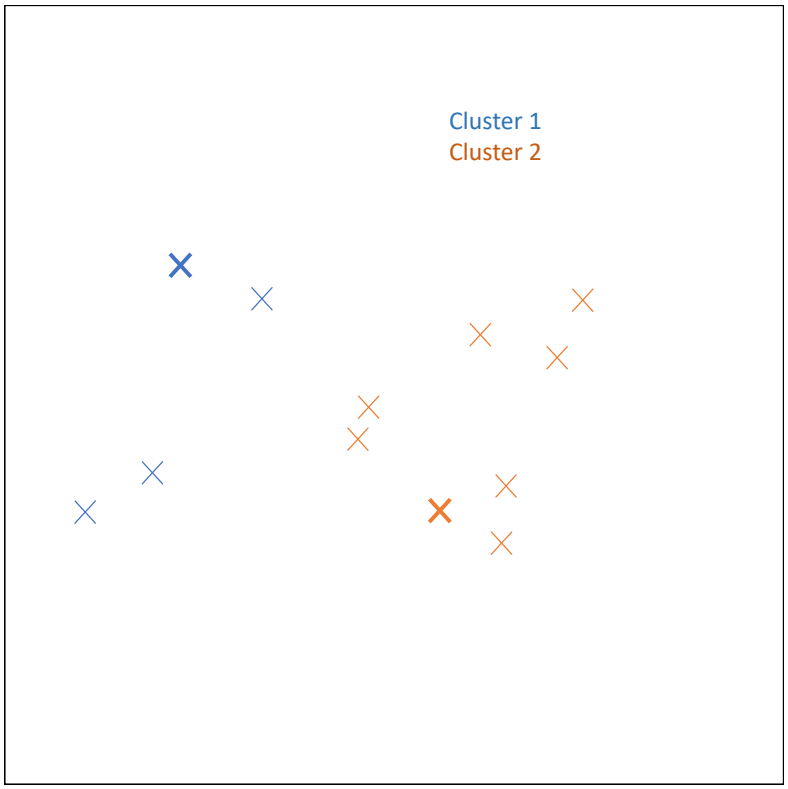


Figure 5.2: Illustration showing the neighbourhoods around two locations (crosses in bold) using the cluster method to define the neighbourhood. A clustering approach selects all observations that have been clustered with each location (blue and orange).

ration of a dependent variable that has been observed at fixed locations at fixed time points, where the locations do not necessarily need to be the same over time. As with the spatial case introduced by Georganos et al. (2019), the methodology can be described as a disaggregation of RF into geographical space and time in the form of local sub models.

The simplest approach for modelling spatio-temporal data using geoRFs would be to assume time independence. Under the assumption of time independence, for each time t we compute a local RF for each location i, including observations that are within the neighbourhood of i. The neighbourhood for each location i at each time t can be defined by fixed or adaptive bandwidths, introduced in Section 5.3. In order to generate predicted values, the predictions from the global forest and local forests are combined in the same way as the spatial case, using Equation 5.1.

However, it is unlikely for spatio-temporal data to be temporally independent. Ideally, the geoRF methodology should explicitly take into account temporal autocorrelation, as it does spatial autocorrelation, in order to improve predictive accuracy. In order to explicitly account for temporal autocorrelation in the geoRF methodology, we propose that the neighbourhood for each location i at time t includes observations from previous time points, according to an autoregressive bandwidth. We define autoregressive fixed bandwidth as follows.

Suppose we have an observation at location i and at time t. A neighbourhood for this observation is defined by an autoregressive fixed bandwidth, where locations are selected to be included in the neighbourhood if they are within a circle of radius d at time t. In addition, locations from previous time points are included in the neighbourhood if they are within a circle of decreasing radius, where the radius shrinks the further back in time it was observed. In other words, locations are included in the neighbourhood of

the observation if they are within a circle of radius,

$$
\begin{aligned}
&d \text{ at time } t,\\
&d\rho \text{ at time } t+1,\\
&d\rho^2 \text{ at time } t+2,\\
&\ldots\\
&d\rho^{T-1} \text{ at time } T,
\end{aligned}
$$

where ρ is a temporal correlation strength parameter. An example of this neighbourhood for an observation at time $t = 1$ is illustrated in the left plot of Figure 5.3.

We also define an autoregressive adaptive bandwidth neighbourhood as one where observations are selected to be included in the neighbourhood of an observation, $\boldsymbol{y}_t(\boldsymbol{s}_i)$, observed at location i, and at time t, when they are the nearest n observations at time t, or the nearest $n\rho$ observations at time $t+1$, or the nearest $n\rho^2$ observations at time $t+2$, and so on, where ρ is a temporal correlation strength parameter. An example of this neighbourhood for an observation at time $t = 1$ is illustrated in the right plot of Figure 5.3.

Similarly, we define an autoregressive adaptive bandwidth as follows. Suppose we have an observation at location i and at time t. A neighbourhood for this observation is defined by an autoregressive adaptive bandwidth, where the nearest n locations are selected to be included in the neighbourhood at time t. In addition, the nearest m locations at a previous time point are included in the neighbourhood, where $m < n$, m decreases the further back in time it was observed. In other words, locations are

included in the neighbourhood of the observation if they the nearest,

$$n \text{ at time } t,$$
$$n\rho \text{ at time } t+1,$$
$$n\rho^2 \text{ at time } t+2,$$
$$\ldots$$
$$n\rho^{T-1} \text{ at time } T,$$

where ρ is a temporal correlation strength parameter. An example of this neighbourhood for an observation at time $t = 1$ is illustrated in the right plot of Figure 5.3.

We also extend our proposed cluster approach to the spatio-temporal case. Temporal autocorrelation was difficult to account for when constructing neighbourhoods using the clustered approach. Therefore, we made the assumption of time-independence for the clustering case only. An illustration of the neighbourhood is given in Figure 5.4

We now perform several simulation experiments to evaluate the predictive accuracy of geoRFs on simulated spatial and spatio-temporal data, using different neighbourhoods.

5.6 Simulation

In this section, we evaluate the performance of RF and several geoRFs on simulated data using different kernel structures. The aim of the simulation experiment is to assess the performance of these approaches in their abilities to accurately predict, and account for autocorrelation, in the values of a dependent variable at new locations, based on a training dataset. In particular, we use two measures of accuracy and a measure of residual spatial autocorrelation over a set of different scenarios. The measures of accuracy that are to be calculated in each scenario are the root mean square error (RMSE), and the mean absolute error (MAE). The measure of residual spatial autocorrelation that is to be calculated in each scenario is

Figure 5.3: Illustration showing the neighbourhoods around two locations at time $t = 1$, using two different methods to define the neighbourhood.

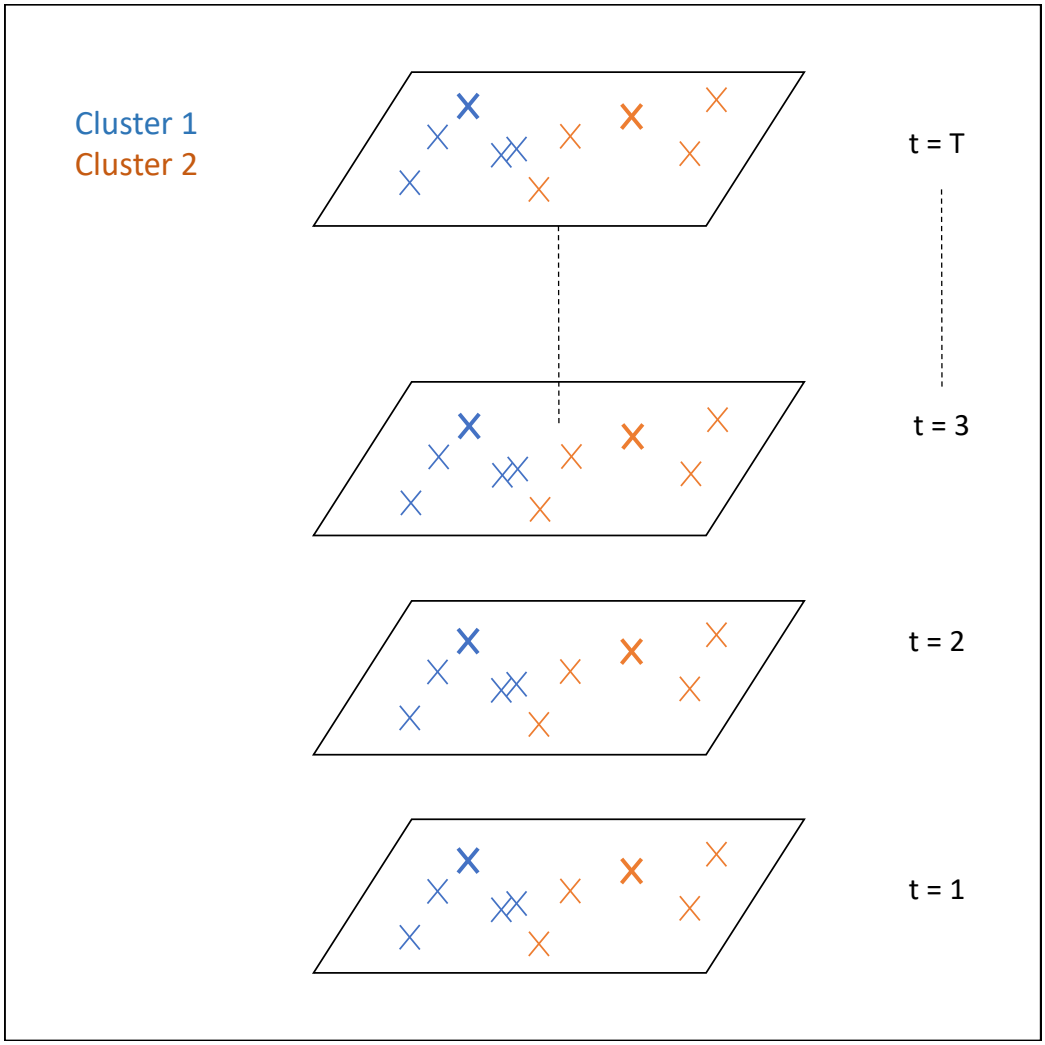


Figure 5.4: Illustration showing the neighbourhoods around two locations at time $t = 1$, using two different methods to define the neighbourhood.

Moran's I, calculated from the residuals. Furthermore, we perform the experiments separately for the spatial case, and the spatio-temporal case. For both cases, the models were determined by varying the neighbourhood structure (fixed, adaptive, or clustered), the bandwidth, and the weight parameter. In the spatio-temporal case, a temporal correlation parameter was also varied. The following sections outline the experimental designs, as well as the simulation procedure.

5.6.1 Spatial simulation

A simulation experiment was conducted to evaluate the performance of RF and geoRFs on simulated data with a spatial structure. The data was generated with motivation from an air pollution context, where a number of stations that measured the concentration of some pollutant were imagined.

We randomly generated $N = 615$ longitude (s_{long}) and latitude (s_{lat}) values from a unit square,

$$s_{\text{long}} \sim \text{U}(0, 1),$$

$$s_{\text{lat}} \sim \text{U}(0, 1).$$

We then simulated five covariates, ensuring that some were spatially correlated and some were not. The inclusion of spatially correlated covariates as features in RF was shown to reduce the amount of residual spatial autocorrelation (Hengl et al., 2018). In the context of the air pollution example, we imagined observing the spatially correlated variables, temperature, windspeed, rainfall, and proximity to a pollutant source. We also imagined observing the spatially uncorrelated variable elevation. The following equations were used to generate these covariates:

$$\text{elevation} : X_1 \sim \text{U}(0, 20),$$

$$\text{temperature} : X_2 = 10 + 2s_{\text{long}} + 10s_{\text{lat}} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 1.5),$$

$$\text{windspeed} : X_3 = 6s_{\text{long}} + 2s_{\text{lat}} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 1),$$

$$\text{rainfall} : X_4 = s_{\text{long}} + s_{\text{lat}} + |\varepsilon|, \varepsilon \sim \text{N}(0, 0.5),$$
$$\text{proximity} : X_5 = \sqrt{(s_{\text{long}} - 0.5)^2 + (s_{\text{lat}} - 0.5)^2}.$$

A dependent variable, $\boldsymbol{y}$, was then simulated from the model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\boldsymbol{X}\boldsymbol{\beta}$ is the linear combination of an intercept and the five covariates, $\boldsymbol{\varepsilon}$ are the errors for the measurement process, and $\boldsymbol{\zeta}$ are the errors for the spatial process, that introduced spatial autocorrelation to $\boldsymbol{y}$. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y} \sim \text{N}(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\boldsymbol{I} + \boldsymbol{\Sigma}), \tag{5.3}$$

where, and $\boldsymbol{\Sigma}$ was the exponential covariance matrix,

$$(\boldsymbol{\Sigma})_{ij} = \sigma^2 \exp\left(\frac{-d_{ij}}{\psi}\right), \tag{5.4}$$

where d_{ij} is the Euclidean distance between location i and j. Here, the parameters $\boldsymbol{\beta}$, τ^2, σ^2, ψ, were all chosen to reflect a possible reality. We set $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 1)'$ so that each covariate had an equal effect on the simulated dependent variable, $\tau^2 = 0.1$ to reduce the influence of the measurement process, $\sigma^2 = 1$ to enhance the presence of spatial autocorrelation within $\boldsymbol{y}$, and $\psi = 0.1$ to induce spatial autocorrelation. We decided that $\sigma^2 > \tau^2$ so that the variability of the measurement process was less than that of the spatial process. The data were sampled using Cholesky factorisation (Algorithm 2, Rue & Held (2005)).

Figure 5.5 displays interpolated surface plots of the five covariates and $\boldsymbol{y}$ that were produced to show the spatial autocorrelation within each of the variables. We see that the dependent variable visually exhibits spatial autocorrelation, with clusters of values that were observed at the upper right region higher compared to those that were observed at the lower left region. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated y values. We calculated $I = 0.1303$, with a

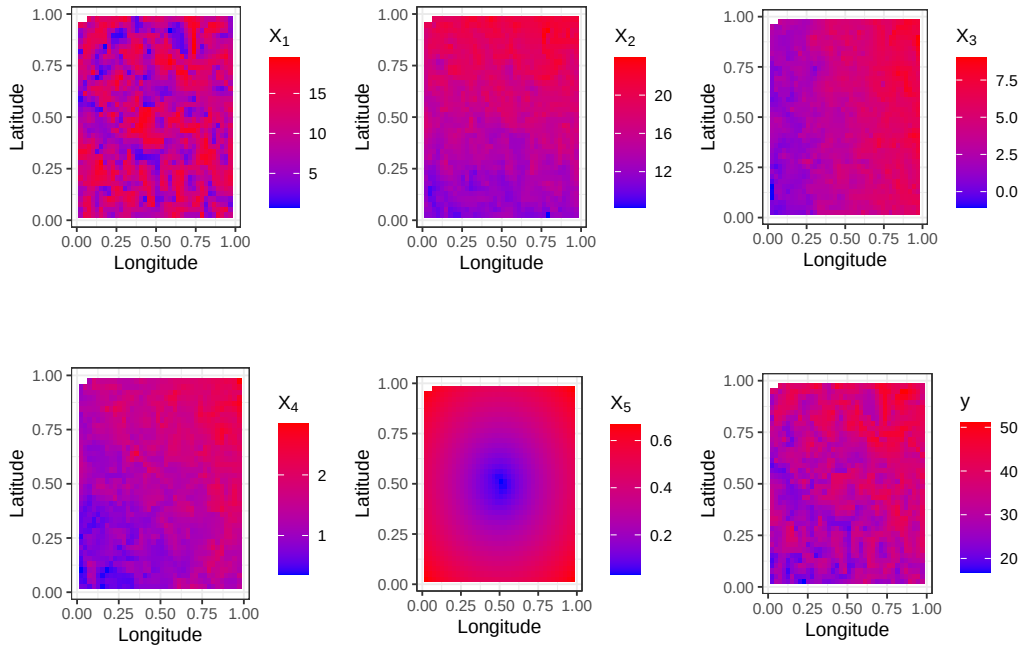


Figure 5.5: Interpolated surface plots of the five randomly generated covariates, and the simulated dependent variable. For X_2 to X_5 spatial autocorrelation is exhibited as expected. Spatial autocorrelation is also evidenced for y. There is a general upward diagonal trend, with higher values displayed at the top right, and lower values displayed at the bottom left.

p-value for the two-sided test for presence of spatial autocorrelation less than 2.2×10^{-16}, confirming the presence of significant spatial autocorrelation within the dependent variable.

We wish to compare geoRF methods that use a fixed neighbourhood structure, adaptive neighbourhood structure, and clustered neighbourhood structure. For each method, the models were determined by varying the bandwidth, and the weight parameter. The experimental design for the methods are given in Table 5.1.

A 10-fold cross validation was performed, where a training set of 500 ob-

Table 5.1: Experimental design for the spatial simulation study.

Kernel	Bandwidth, b	Weight parameter, a	Total
Adaptive	100, 200, 300, 400	0, 0.25, 0.5, 0.75, 1	$4 \times 5 = 20$
Fixed	0.2, 0.35, 0.5, 0.65	0, 0.25, 0.5, 0.75, 1	$4 \times 5 = 20$
Cluster	2, 3, 4, 5	0, 0.25, 0.5, 0.75, 1	$4 \times 5 = 20$
Total			60 models

servations was randomly sampled from the 615 simulated observations, without replacement, 10 times. Each time, the remaining 115 observations were put aside as the test set. We fit the 60 models to each of the training sets to train the models, and the test sets were used to compute predictions at new locations and to calculate the RMSE, MAE, and Moran's I on the residuals. The mean RMSE, and mean MAE were computed to compare the performance of each model.

Figure 5.6 displays the mean RMSE, mean MAE, and mean Moran's I calculated over each test set, for each model. When an adaptive kernel (left three plots in Figure 5.6) was used to select the observations to be included in the neighbourhood of each local RF for each location in the training set, the weight parameter that resulted in the lowest mean RMSE and lowest mean MAE was $a = 0$, for all bandwidths tried. A weight parameter of $a = 0$ corresponds to the traditional RF approach. As the weight parameter increased towards 1, the mean RMSE and mean MAE generally increased. Furthermore, as the bandwidth was increased to include more observations in each local RF, the mean RMSE and mean MAE increased less with increasing a. This is because when more observations are included in the local RF for each observation, the more accurate the predictions become. There appeared to be little to no change in mean RMSE and mean MAE as a was increased when the adaptive bandwidth was 300 or greater. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the adaptive kernel.

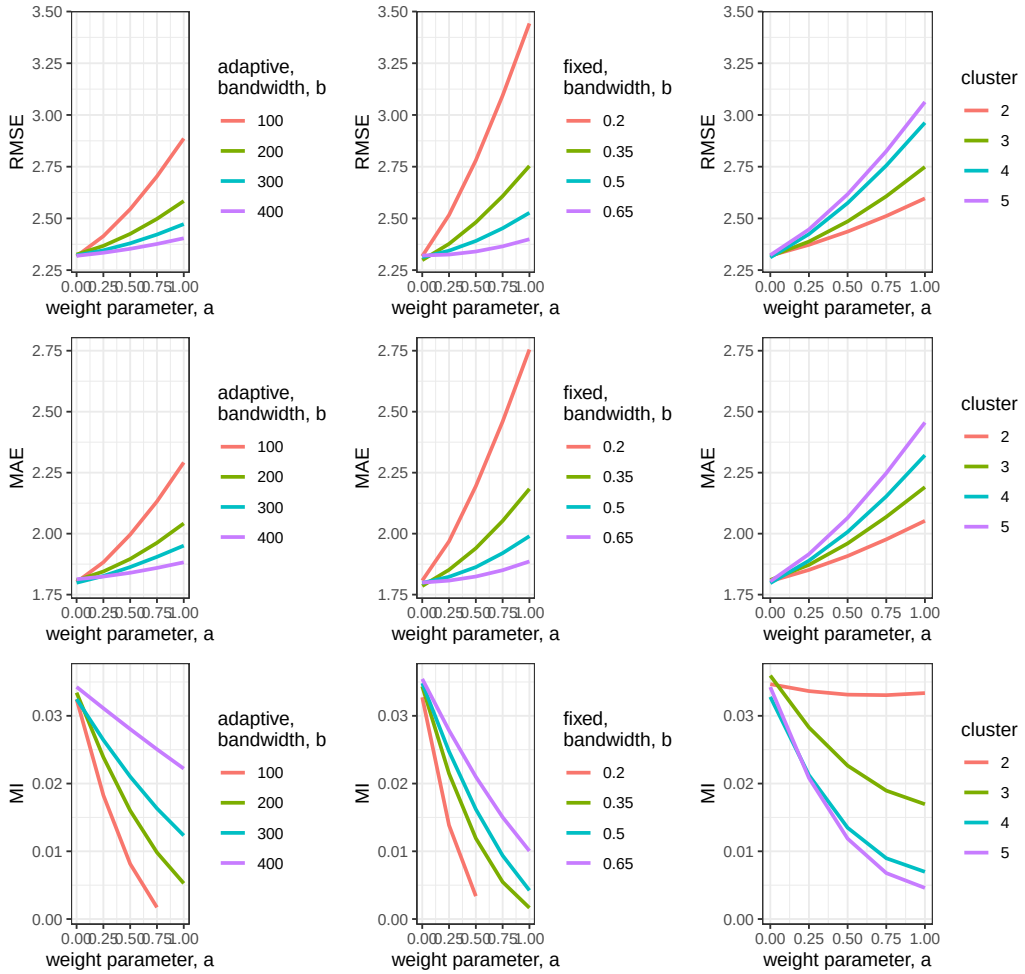


Figure 5.6: Mean measures of accuracy and spatial autocorrelation for each approach. On the left, an adaptive kernel approach was used for the geoRF, with bandwidths 100, 200, 300, and 400 tried. In the middle, a fixed kernel approach was used for the geoRF, with bandwidths 0.4, 0.6, 0.8, and 1 tried. On the right, our clustering approach was used for the geoRF, with 2, 3, 4, and 5 clusters tried. The weighting parameter is displayed on the x-axis.

Including predictions calculated from local RF on each observation did not appear to increase the predictive accuracy for new data, irrespective of bandwidth and weight parameter when an adaptive kernel was used. However, when the weight parameter increased from $a = 0$ to $a = 1$ the mean absolute value of Moran's I decreased. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all bandwidths tried. Furthermore, as the bandwidth was decreased to include less observations in each local RF, the mean absolute value of Moran's I decreased more with increasing a. Including more predictions from local RF on each observation decreases the amount of residual spatial autocorrelation. This suggests that the local RF at each location were able to take into account the spatial autocorrelation within the data.

When a fixed kernel (middle three plots in Figure 5.6) was used to select the observations to be included to be included in the neighbourhood of each local RF for each location in the training set, the weight parameter that resulted in the lowest mean RMSE and mean MAE was $a = 0$ for all bandwidths tried, corresponding to the traditional RF approach. This was the same trend observed in the adaptive kernel case. As the weight parameter increased towards 1, the mean RMSE and mean MAE increased for most bandwidths tried. In general, as the fixed bandwidth increased to cover more locations and include them in the local RF for each observation, the more accurate the predictions become. There appeared to be little to no change in mean RMSE as a was increased when the fixed bandwidth was 065. Further, there appeared to be little to no change in mean MAE as a increased when the fixed bandwidth was 0.8. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the fixed kernel.

Once again, including predictions calculated from local RF on each observation did not appear to increase the predictive accuracy for new data, irrespective of bandwidth and weight parameter when a fixed kernel was used. However, when the weight parameter increased from $a = 0$ to $a = 1$

the mean absolute value of Moran's I decreased. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all bandwidths tried. Furthermore, as the bandwidth was decreased to cover less locations and therefore less observations in each local RF, the mean absolute value of Moran's I decreased more with increasing a. Including more predictions from local RF on each observation decreases the amount of residual spatial autocorrelation. This suggests that the local RF at each location were able to take into account the spatial autocorrelation within the data.

When a clustering approach was taken to selecting observations to be included in the neighbourhood of each local RF for each cluster in the training set, the weight parameter that resulted in the lowest mean RMSE and lowest mean MAE was $a = 0$ for every scenario with a different number of clusters tried. Once again, this corresponds to the traditional RF approach, and was the same trend observed for both the adaptive and fixed kernel cases. As the weight parameter increased towards 1, the mean RMSE and mean MAE increased for each scenario with a different number of clusters tried. In general, as the number of clusters increased, the number of observations within each cluster, and hence, within the neighbourhood of the local RFs, decreased. This meant that we observed less accurate predictions as the number of clusters increased. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the fixed kernel.

Once more, including predictions calculated from local RFs on each cluster did not appear to increase the predictive accuracy for new data, irrespective of the number of clusters and weight parameter. However, when the weight parameter increased from $a = 0$ to $a = 1$ the mean absolute value of Moran's I decreased. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all numbers of clusters tried. Furthermore, as the number of clusters was increased, the mean absolute value of Moran's I decreased more with increasing a. Including

predictions from local RFs from more clusters decreases the amount of residual spatial autocorrelation. This suggests that the local RF for each cluster were able to take into account the spatial autocorrelation within the data.

5.6.2 Spatio-temporal simulation

We generalized the geoRF approach above to the spatio-temporal case. We assumed that both spatial and temporal autocorrelation existed within a continuous response variable observed at fixed locations in (2D) space and time, and that the locations did not change over time. Further, we assumed that there was no interaction between space and time.
A simulation experiment was conducted to evaluate the performance of RF and geoRFs on simulated data with a spatio-temporal structure. The data was generated with motivation from an air pollution context, where a number of stations that measured the concentration of some pollutant over time were imagined. Similar to the spatial simulation, we randomly generated $N = 313$ longitude (s_{long}) and latitude (s_{lat}) values from a unit square,

$$s_{\text{long}} \sim \text{U}(0, 1),$$
$$s_{\text{lat}} \sim \text{U}(0, 1).$$

We then simulated five covariates for $T = 5$ time points, ensuring that some were spatially correlated and some were not. In the context of the air pollution example, we imagined observing the spatially correlated variables, temperature, windspeed, rainfall, proximity to a pollutant source. We also imagined observing the spatially uncorrelated variable elevation. The following equations were used to generate these covariates:

$$\text{elevation} : X_{1t} \sim \text{U}(0, 20),$$
$$\text{temperature} : X_{2t} = 10 + 2s_{\text{long}} + 10s_{\text{lat}} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 1.5),$$

Table 5.2: 10-fold cross validated RMSE, MAE, and MI for the spatial simulation

Kernel	b	a	RMSE	MAE	MI	Kernel	b	a	RMSE	MAE	MI
Adaptive	100	0	2.48	1.95	0.036	Fixed	0.4	0	2.46	1.93	0.038
Adaptive	100	0.25	2.56	2.03	0.019	Fixed	0.4	0.25	2.65	2.10	0.017
Adaptive	100	0.5	2.68	2.15	0.006	Fixed	0.4	0.5	2.90	2.32	0.003
Adaptive	100	0.75	2.83	2.30	-0.003	Fixed	0.4	0.75	3.20	2.57	-0.003
Adaptive	100	1	3.00	2.45	-0.008	Fixed	0.4	1	3.53	2.86	-0.005
Adaptive	200	0	2.48	1.95	0.039	Fixed	0.6	0	2.45	1.93	0.040
Adaptive	200	0.25	2.51	1.98	0.027	Fixed	0.6	0.25	2.52	1.99	0.024
Adaptive	200	0.5	2.55	2.02	0.017	Fixed	0.6	0.5	2.61	2.07	0.012
Adaptive	200	0.75	2.61	2.08	0.009	Fixed	0.6	0.75	2.72	2.17	0.003
Adaptive	200	1	2.68	2.15	0.003	Fixed	0.6	1	2.87	2.28	-0.002
Adaptive	300	0	2.45	1.93	0.040	Fixed	0.8	0	2.46	1.94	0.040
Adaptive	300	0.25	2.47	1.95	0.031	Fixed	0.8	0.25	2.48	1.96	0.029
Adaptive	300	0.5	2.50	1.98	0.024	Fixed	0.8	0.5	2.52	2.00	0.020
Adaptive	300	0.75	2.54	2.02	0.017	Fixed	0.8	0.75	2.57	2.05	0.012
Adaptive	300	1	2.58	2.06	0.011	Fixed	0.8	1	2.63	2.10	0.006
Adaptive	400	0	2.48	1.95	0.037	Fixed	1	0	2.47	1.94	0.040
Adaptive	400	0.25	2.48	1.96	0.034	Fixed	1	0.25	2.47	1.95	0.032
Adaptive	400	0.5	2.50	1.97	0.032	Fixed	1	0 .5	2.48	1.96	0.025
Adaptive	400	0.75	2.51	1.99	0.030	Fixed	1	0.75	2.50	1.98	0.019
Adaptive	400	1	2.53	2.01	0.028	Fixed	1	1	2.53	2. 01	0.014
Cluster	2	0	2.49	1.97	0.039	Cluster	4	0	2.48	1.94	0.37
Cluster	2	0.25	2.52	2.00	0.032	Cluster	4	0.25	2.60	2.05	0.025
Cluster	2	0.5	2.57	2.04	0.027	Cluster	4	0.5	2.75	2.18	0.016
Cluster	2	0.75	2.64	2.10	0.022	Cluster	4	0.75	2.94	2.33	0.010
Cluster	2	1	2.71	2.16	0.018	Cluster	4	1	3.15	2.50	0.007
Cluster	3	0	2.48	1.94	0.035	Cluster	5	0	2.47	1.94	0.039
Cluster	3	0.25	2.53	2.01	0.026	Cluster	5	0.25	2.60	2.06	0.023
Cluster	3	0.5	2.62	2.10	0.019	Cluster	5	0.5	2.77	2.22	0.012
Cluster	3	0.75	2.73	2.20	0.014	Cluster	5	0.75	2.99	2.40	0.005
Cluster	3	1	2.87	2.33	0.011	Cluster	5	1	3.25	2.61	0.002

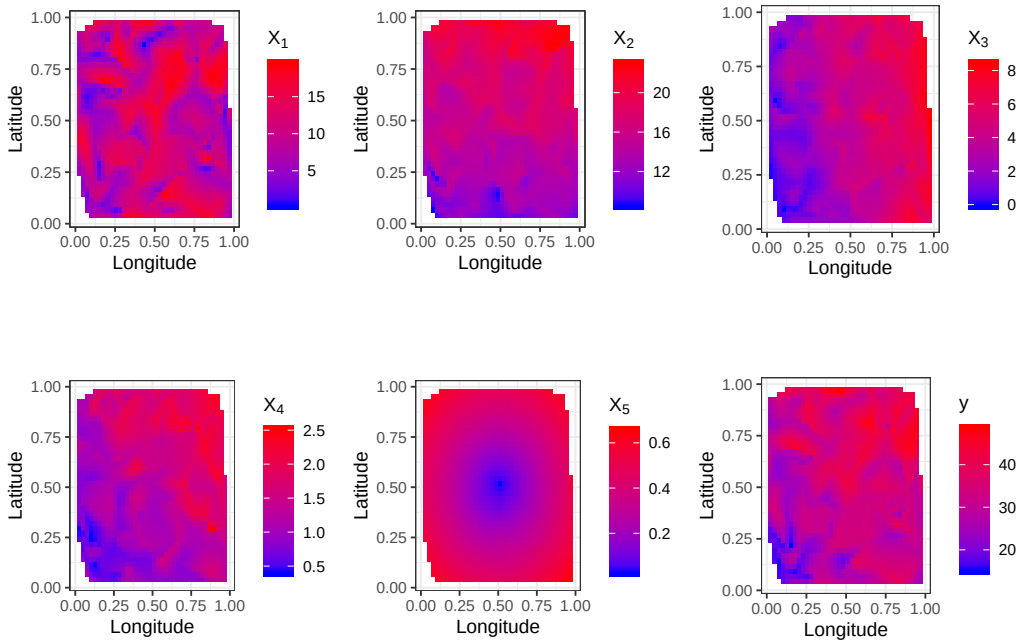


Figure 5.7: Interpolated surface plots of the five randomly generated covariates for time point $t = 1$, and the simulated dependent variable. For X_2 to X_5 spatial autocorrelation is exhibited as expected. Spatial autocorrelation is also evidenced for y. There is a general upward diagonal trend, with higher values displayed at the top right, and lower values displayed at the bottom left. Similar patterns were observed for $t = 2, 3, 4, 5$.

$$\text{windspeed} : X_{3t} = 6s_{\text{long}} + 2s_{\text{lat}} + \varepsilon, \quad \varepsilon \sim \text{N}(0, 1),$$
$$\text{rainfall} : X_{4t} = s_{\text{long}} + s_{\text{lat}} + |\varepsilon|, \varepsilon \sim \text{N}(0, 0.5),$$
$$\text{proximity} : X_{5t} = \sqrt{(s_{\text{long}} - 0.5)^2 + (s_{\text{lat}} - 0.5)^2}.$$

A dependent variable, $\boldsymbol{y}_t$, was then simulated from the model,

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\zeta}_t + \boldsymbol{\varepsilon}_t, \tag{5.5}$$

where $\boldsymbol{X}_t\boldsymbol{\beta}$ is the linear combination of an intercept and the five covariates observed at time t, $\boldsymbol{\varepsilon}_t$ are the errors for the measurement process observed

at time t, and $\boldsymbol{\zeta}_t$ are the errors for the spatial process observed at time t, that introduced spatial autocorrelation to $\boldsymbol{y}_t$. Explicitly, the dependent variable was drawn from,

$$\boldsymbol{y}_t \sim \mathrm{N}(\boldsymbol{X}_t\boldsymbol{\beta}, \tau^2\boldsymbol{I} + \boldsymbol{\Sigma}), \tag{5.6}$$

where, and $\boldsymbol{\Sigma}$ was the exponential covariance matrix,

$$(\boldsymbol{\Sigma})_{ij} = \sigma^2 \exp\left(\frac{-d_{ij}}{\psi}\right), \tag{5.7}$$

where d_{ij} is the Euclidean distance between location i and j. Here, the parameters $\boldsymbol{\beta}$, τ^2, σ^2, ψ, were all chosen to reflect a possible reality. We set $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 1)'$, so that each covariate had an equal effect on the simulated dependent variable, $\tau^2 = 0.1$ to reduce the influence of the measurement process, $\sigma^2 = 1$ to enhance the presence of spatial autocorrelation within $\boldsymbol{y}_t$, and $\psi = 0.1$ to induce spatial autocorrelation. We decided that $\sigma^2 > \tau^2$ so that the variability of the measurement process was less than that of the spatial process. The data was sampled using Cholesky factorisation (Algorithm 2, Rue & Held (2005)).

Figure 5.7 displays interpolated surface plots of the five covariates and the dependent variable for $t = 1$ that were produced to show the spatial autocorrelation within each of the variables. We see that the dependent variable visually exhibits spatial autocorrelation, with clusters of values that were observed at the upper right region higher compared to those that were observed at the lower left region. This trend was observed for the rest of the time points, $t = 2, 3, 4, 5$. Moran's I was calculated to confirm the presence of spatial autocorrelation in the simulated $\boldsymbol{y}_t$ values for each t. We calculated $I_{t=1} = 0.1436, I_{t=2} = 0.1462, I_{t=3} = 0.1279, I_{t=4} = 0.1176,$ and $I_{t=5} = 0.1120$, all with corresponding p-values for the two-sided tests for presence of spatial autocorrelation less than 2.2×10^{-16}, confirming the presence of significant spatial autocorrelation within the dependent variable.

We wish to compare geoRF methods that use a fixed neighbourhood structure, adaptive neighbourhood structure, and clustered neighbourhood struc-

Table 5.3: Experimental design for the spatio-temporal simulation study.

Kernel	Bandwidth, b	Weight parameter, a	Temporal correlation, ρ	Total
Adaptive	30, 60, 90, 120	0, 0.5, 1	0, 0.33, 0.67, 1	$4 \times 3 \times 4 = 48$
Fixed	0.2, 0.35, 0.5, 0.65	0, 0.5, 1	0, 0.33, 0.67, 1	$4 \times 3 \times 4 = 48$
Cluster	2, 3, 4, 5	0, 0.5, 1	NA	$4 \times 3 = 12$
Total				108 models

ture. For each method, the models were determined by varying the bandwidth, the weight parameter, and the correlation parameter, ρ. The experimental design for the methods are given in Table 5.1.

For each model, all covariates were included and eligible to be selected as features in the construction of the forests. Furthermore, the variable "time" (taking values $t = 1, ..., 5$) was included, because it was found to be the most important factor for prediction according to the spatio-temporal RF of Hengl et al. (2018). Furthermore, since time was considered important, the number of features to be selected at each split was chosen to be 4, rather than the default, to ensure that time is almost always chosen as a feature.

A 10-fold cross validation was performed, where a training set of 250 observations was randomly sampled from the 313 simulated observations, without replacement, 10 times. Each time, the remaining 63 observations were put aside as the test set. We fit the 108 models to each of the training sets to train the models, and the test sets were used to compute predictions at new locations and to calculate the RMSE, MAE, and Moran's I on the residuals. The mean RMSE, and mean MAE were computed to compare the performance of each model.

Figure 5.8 displays the mean RMSE and mean MAE, and mean Moran's I for time $t = 1$, each calculated over all test sets, for each model. When an adaptive kernel (left three plots in Figure 5.8) was used to select the observations to be included in the neighbourhood of each local RF for

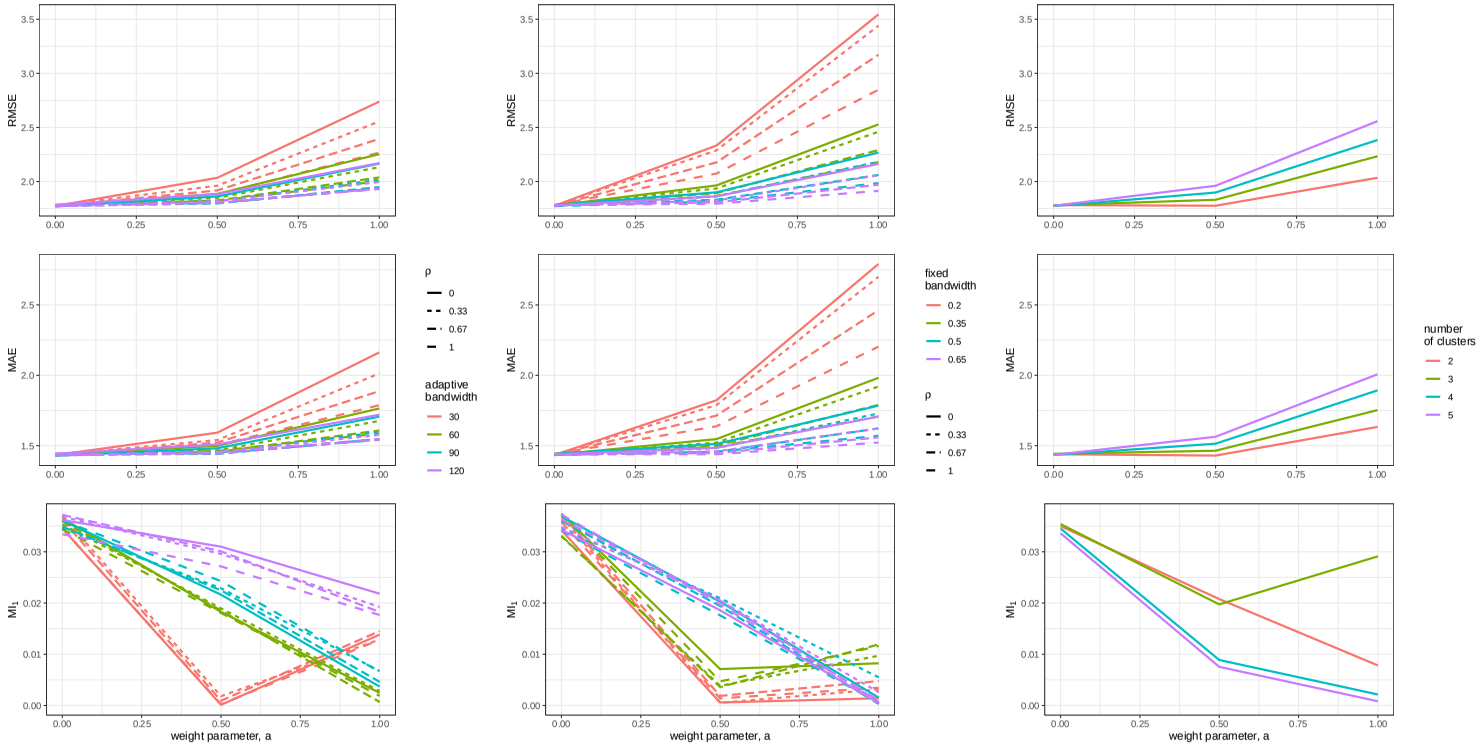


Figure 5.8: Mean measures of accuracy, and mean measures of spatial autocorrelation for time $t = 1$, for each approach. On the left, an adaptive kernel approach was used for the geoRF, with bandwidths 50, 100, 150, and 200 tried. In the middle, a fixed kernel approach was used for the geoRF, with bandwidths 0.4, 0.6, 0.8, and 1 tried. On the right, our clustering approach was used for the geoRF, with 2, 3, 4, and 5 clusters tried. The line type represents the temporal correlation parameters tried. The weighting parameter is displayed on the x-axis.

each observation in the training set, the weight parameter that resulted in the lowest mean RMSE and lowest mean MAE was $a = 0$, for all bandwidths and temporal correlation parameters tried. A weight parameter of $a = 0$ corresponds to the traditional RF approach, that includes all observations, over all time points. As the weight parameter increased towards 1, the mean RMSE and mean MAE generally increased. Furthermore, as the bandwidth was increased to include more observations in each local RF, the mean RMSE and mean MAE increased more slowly with increasing a. This is because when more observations were included in the local RF for each observation, the more accurate the predictions become. Furthermore, when the temporal correlation parameter was increased from 0 to 1, mean RMSE and mean MAE increased for all bandwidths tried, and for all weight parameters tried. This was due to the fact that increasing the temporal correlation parameter again causes more observations to be included in the neighbourhood of each observation and therefore each local RF produces more accurate predictions. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the adaptive kernel.

Including predictions calculated from local RF on each observation did not appear to increase the predictive accuracy for new data, irrespective of bandwidth and correlation parameter when an adaptive kernel was used. However, when the weight parameter increased from $a = 0$ to $a = 1$ the mean absolute value of Moran's I when $t = 1$ decreased overall. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all bandwidths and correlation parameters tried. The weight parameter $a = 0.5$ displayed similar results. Furthermore, as the bandwidth was decreased to include less observations in each local RF, the mean absolute value of Moran's I decreased. This suggests that the local RF at each location were able to take into account the spatial autocorrelation within the data.

When a fixed kernel (middle three plots in Figure 5.8) was used to select

the observations to be included to be included in the neighbourhood of each local RF for each observation in the training set, the weight parameter that resulted in the lowest mean RMSE and lowest mean MAE was $a = 0$ for all bandwidths and temporal correlation parameters tried. This corresponds to the traditional RF approach, that includes all observations over all time points. This was the same trend observed in the adaptive kernel case. As the weight parameter increased towards 1, the mean RMSE and mean MAE generally increased. Furthermore, as the fixed bandwidth was increased to cover and include more observations in the local RF for each observation, the more accurate the predictions became. Further still, when the correlation parameter was increased from 0 to 1, mean RMSE and mean MAE increased for all bandwidths tried, and for all weight parameters tried. This was again due to the fact that increasing the temporal correlation parameter caused more observations to be included in the neighbourhood of each observation and therefore each local RF produced more accurate predictions. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the fixed kernel.

Once again, including predictions calculated from local RF on each observation did not appear to increase the predictive accuracy for new data, irrespective of bandwidth and correlation parameter when a fixed kernel was used. However, when the weight parameter increased from $a = 0$ to $a = 1$ the mean absolute value of Moran's I when $t = 1$ decreased overall. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all bandwidths and correlation parameters tried. The weight parameter $a = 0.5$ displayed similar results. Furthermore, as the bandwidth was decreased to include less observations in each local RF, the mean absolute value of Moran's I decreased. This suggested that the local RF at each location were able to take into account the spatial autocorrelation within the data.

When a clustering approach was taken to selecting observations to be in-

cluded in the neighbourhood of each local RF for each cluster in the training set, the weight parameter that resulted in the lowest mean RMSE and lowest mean MAE was $a = 0$ for every scenario with a different number of clusters tried. Once again, this corresponds to the traditional RF approach, and was the same trend observed for both the adaptive and fixed kernel cases. As the weight parameter increased towards 1, the mean RMSE and mean MAE increased for each scenario with a different number of clusters tried. In general, as the number of clusters increased, the number of observations within each cluster, and hence, within the neighbourhood of the local RFs, decreased. This meant that we observed less accurate predictions in general as the number of clusters increased. However, when the number of clusters was small, the mean RMSE and mean MAE did not increase by very much when the weight parameter was increased. Overall, the traditional RF produced the most accurate predictions on new data, according to RMSE and MAE for the fixed kernel.

Once more, including predictions calculated from local RFs on each cluster did not appear to increase the predictive accuracy for new data, irrespective of the number of clusters and weight parameter. However, when the weight parameter increased from $a = 0$ to $a = 1$ the mean absolute value of Moran's I decreased. The weight parameter that resulted in the lowest mean absolute value for Moran's I was $a = 1$ for all numbers of clusters tried. Furthermore, as the number of clusters was increased, the mean absolute value of Moran's I decreased more with increasing a. Including predictions from local RFs from more clusters decreases the amount of residual spatial autocorrelation. This suggests that the local RF for each cluster were able to take into account the spatial autocorrelation within the data.

5.6.3 Summary

The clustering geoRF method did not produce as or more accurate predictions compared to the traditional RF, in both the spatial and spatio-

Table 5.4: 10-fold cross validated RMSE, MAE for the spatio-temporal simulation

Kernel	b	ρ	a	RMSE	MAE
Adaptive	20	0	0	1.91	1.52
Adaptive	20	0.33	0	1.92	1.53
Adaptive	20	0.67	0	1.92	1.53
Adaptive	20	1	0	1.92	1.53
Adaptive	20	0	0.5	2.32	1.79
Adaptive	20	0.33	0.5	2.24	1.73
Adaptive	20	0.67	0.5	2.14	1.66
Adaptive	20	1	0.5	2.11	1.63
Adaptive	20	0	1	3.15	2.41
Adaptive	20	0.33	1	2.94	2.27
Adaptive	20	0.67	1	2.70	2.06
Adaptive	20	1	1	2.57	1.96
Adaptive	40	0	0	1.92	1.53
Adaptive	40	0.33	0	1.92	1.53
Adaptive	40	0.67	0	1.91	1.52
Adaptive	40	1	0	1.91	1.52
Adaptive	40	0	0.5	2.06	1.63
Adaptive	40	0.33	0.5	2.00	1.58
Adaptive	40	0.67	0.5	1.98	1.57
Adaptive	40	1	0.5	1.98	1.56
Adaptive	40	0	1	2.47	1.92
Adaptive	40	0.33	1	2.27	1.77
Adaptive	40	0.67	1	2.20	1.73
Adaptive	40	1	1	2.17	1.70

Kernel	b	ρ	a	RMSE	MAE
Adaptive	60	0	0	1.92	1.53
Adaptive	60	0.33	0	1.92	1.53
Adaptive	60	0.67	0	1.91	1.52
Adaptive	60	1	0	1.91	1.52
Adaptive	60	0	0.5	2.03	1.62
Adaptive	60	0.33	0.5	1.99	1.58
Adaptive	60	0.67	0.5	1.98	1.57
Adaptive	60	1	0.5	1.98	1.56
Adaptive	60	0	1	2.38	1.88
Adaptive	60	0.33	1	2.21	1.74
Adaptive	60	0.67	1	2.20	1.73
Adaptive	60	1	1	2.17	1.70
Adaptive	80	0	0	1.92	1.53
Adaptive	80	0.33	0	1.92	1.53
Adaptive	80	0.67	0	1.91	1.52
Adaptive	80	1	0	1.92	1.53
Adaptive	80	0	0.5	2.03	1.62
Adaptive	80	0.33	0.5	1.99	1.58
Adaptive	80	0.67	0.5	1.98	1.57
Adaptive	80	1	0.5	1.98	1.57
Adaptive	80	0	1	2.38	1.88
Adaptive	80	0.33	1	2.21	1.74
Adaptive	80	0.67	1	2.15	1.71
Adaptive	80	1	1	2.14	1.68

Table 5.5: 10-fold cross validated RMSE, and MAE for the spatio-temporal simulation

Kernel	b	ρ	a	RMSE	MAE
Fixed	0.2	0	0	1.91	1.52
Fixed	0.2	0.33	0	1.92	1.53
Fixed	0.2	0.67	0	1.92	1.53
Fixed	0.2	1	0	1.93	1.53
Fixed	0.2	0	0.5	2.74	2.09
Fixed	0.2	0.33	0.5	2.69	2.06
Fixed	0.2	0.67	0.5	2.55	1.96
Fixed	0.2	1	0.5	2.36	1.82
Fixed	0.2	0	1	4.26	3.19
Fixed	0.2	0.33	1	4.15	3.11
Fixed	0.2	0.67	1	3.81	2.85
Fixed	0.2	1	1	3.36	2.51
Fixed	0.35	0	0	1.91	1.52
Fixed	0.35	0.33	0	1.92	1.53
Fixed	0.35	0.67	0	1.92	1.53
Fixed	0.35	1	0	1.91	1.52
Fixed	0.35	0	0.5	2.22	1.73
Fixed	0.35	0.33	0.5	2.18	1.70
Fixed	0.35	0.67	0.5	2.12	1.65
Fixed	0.35	1	0.5	2.04	1.59
Fixed	0.35	0	1	2.91	2.21
Fixed	0.35	0.33	1	2.81	2.13
Fixed	0.35	0.67	1	2.61	1.98
Fixed	0.35	1	1	2.39	1.81

Kernel	b	ρ	a	RMSE	MAE
Fixed	0.5	0	0	1.92	1.53
Fixed	0.5	0.33	0	1.92	1.53
Fixed	0.5	0.67	0	1.91	1.53
Fixed	0.5	1	0	1.92	1.53
Fixed	0.5	0	0.5	2.07	1.63
Fixed	0.5	0.33	0.5	2.03	1.59
Fixed	0.5	0.67	0.5	1.98	1.57
Fixed	0.5	1	0.5	1.98	1.56
Fixed	0.5	0	1	2.50	1.97
Fixed	0.5	0.33	1	2.39	1.85
Fixed	0.5	0.67	1	2.24	1.75
Fixed	0.5	1	1	2.17	1.69
Fixed	0.65	0	0	1.92	1.53
Fixed	0.65	0.33	0	1.92	1.53
Fixed	0.65	0.67	0	1.92	1.53
Fixed	0.65	1	0	1.92	1.53
Fixed	0.65	0	0.5	2.06	1.63
Fixed	0.65	0.33	0.5	2.03	1.60
Fixed	0.65	0.67	0.5	1.97	1.56
Fixed	0.65	1	0.5	1.98	1.57
Fixed	0.65	0	1	2.45	1.91
Fixed	0.65	0.33	1	2.34	1.83
Fixed	0.65	0.67	1	2.17	1.71
Fixed	0.65	1	1	2.15	1.69

Table 5.6: 10-fold cross validated RMSE, MAE for the spatio-temporal simulation

Kernel	k	a	RMSE	MAE
Cluster	2	0	1.91	1.52
Cluster	2	0.5	2.03	1.60
Cluster	2	1	2.30	1.79
Cluster	3	0	1.92	1.53
Cluster	3	0.5	2.03	1.59
Cluster	3	1	2.43	1.90

Kernel	k	a	RMSE	MAE
Cluster	4	0	1.92	1.53
Cluster	4	0.5	2.18	1.71
Cluster	4	1	2.73	2.10
Cluster	5	0	1.92	1.52
Cluster	5	0.5	2.17	1.70
Cluster	5	1	2.77	2.14

temporal simulation studies, with any weight parameter greater than 0, and for any number of clusters. This was thought to be because the cluster approach selected similar, if not the same, locations to be include in the neighbourhoods for each observation as the adaptive bandwidth approach. Essentially, the cluster approach meant that locations were included in the neighbourhood of an observation if they were close to the cluster center for the cluster that the observation belonged to. As the number of clusters increased, fewer locations were included in the neighbourhoods for each observation, and the observations were closer to their cluster centers. Therefore, selecting locations based on clusters essentially became selecting the nearest n observations. The cluster method did account for more spatial autocorrelation than the traditional RF, as evidenced by a decrease in Moran's I on the residuals.

From the spatial simulation study, we have observed a trade off between accuracy and accounting for spatial autocorrelation. It was shown in Figure 5.6 that increasing the bandwidth in the adaptive and fixed models increased the accuracy more as the weight parameter increased. However, this in-turn increased the amount of residual spatial autocorrelation. A "best model" would be one that is as accurate as possible at making predictions on new data while still accounting for spatial autocorrelation. In the spatial simulation, when an adaptive bandwidth of 300, or a fixed bandwidth of 0.65 was used, we saw very little increase in mean RMSE and mean MAE when the weight parameter was increased from 0 to 1, compared to other models of smaller bandwidths. However, we saw a striking decrease in the amount of residual spatial autocorrelation when the same bandwidths were used. A similar trend was also observed for the cluster models. When the number of clusters used was 2, there was little increase in mean RMSE and mean MAE compared to models that used a larger number of clusters to form the neighbourhoods when the weight parameter was increased from 0 to 1. Again, we observed a corresponding decrease in the amount of residual spatial autocorrelation. We

therefore concluded that the best models were the ones with an adaptive bandwidth of 300, a fixed bandwidth of 0.65, and two clusters, when the weight parameter was set to 1. Their accuracy were on par with the traditional RFs but accounted for more spatial autocorrelation.

From the spatio-temporal simulation study, we also observed a trade off between accuracy and accounting for spatial autocorrelation. It was shown in Figure 5.8 that increasing the bandwidth in the adaptive and fixed models increased the accuracy more as the weight parameter increased. Further, increasing the temporal correlation parameter also increased the accuracy more as the weight parameter increased. However, this in-turn increased the amount of residual spatial autocorrelation. Like the spatial case, a "best model" would be one that is as accurate as possible at making predictions on new data while still accounting for spatial autocorrelation. In the spatio-temporal simulation, when an adaptive bandwidth of 120, or a fixed bandwidth of 0.65 was used, we saw very little increase in mean RMSE and mean MAE when the weight parameter was increased from 0 to 1, compared to other models of smaller bandwidths. However, we saw a striking decrease in the amount of residual spatial autocorrelation when the same bandwidths were used. A similar trend was also observed for the cluster models. When the number of clusters used was 2, there was little increase in mean RMSE and mean MAE compared to models that used a larger number of clusters to form the neighbourhoods when the weight parameter was increased from 0 to 1. Again, we observed a corresponding decrease in the amount of residual spatial autocorrelation. We therefore concluded that the best models were the ones with an adaptive bandwidth of 120, a fixed bandwidth of 0.65, and two clusters, when the weight parameter was set to 1, and when the correlation parameter was set to 1. Their accuracy were on par with the traditional RFs but accounted for more spatial autocorrelation.

In these simulation experiments, we compared geoRFs using two sets of simulated data, one for the spatial case, and one for the spatio-temporal

case. In each setting, the data were sampled using Gaussian processes with exponential covariance functions to induce spatial variation. Further research is required in order to study the application of geoRF methods using simulated data generated from different covariance functions. We hypothesize that geoRF would be able to account for spatial autocorrelation, regardless of the spatial structure of the data.

Due to computational limitations and time restraints, the spatial simulation study was performed using a training data set with $N = 500$ locations, and the spatio-temporal simulation study was performed using a training data set with $N = 250$ locations and $T = 5$ time points. Ideally, more locations and time points would have been better. However, the spatial simulation was performed using less locations, and provided a similar trend compared to the $N = 500$ case (albeit the predictions were less accurate). We conclude that, although it would have been better to perform simulations with $N = 10,000$ locations and $T = 100$ time points, the studies performed are adequate in comparing the geoRF methodologies for this thesis.

We now present the application of the geoRF methodology to two real data sets. We first present the results from computing spatial geoRFs on New Zealand particulate matter data for the year 2013, in order to predict particulate matter concentration at unobserved locations. We then present the results from computing spatio-temporal geoRFs on sub-Antarctic hoki catch weight data for the years 2000 – 2008, in order to produce a prediction map for each year.

5.7 Case studies

5.7.1 New Zealand particulate matter

We performed a case study, in which we computed RF and geoRF on the New Zealand particulate matter (PM10) concentration data described in Section 2.7.1. The aim was to find the best forest to predict PM10 concentration across New Zealand, in terms of predictive accuracy RMSE, and MAE, while also accounting for spatial autocorrelation. Once a suitable forest is chosen, we use it to produce an interpolated predictive map for particulate matter concentration, using covariate observations where particulate matter was not observed. Limited covariates were available to compute the RF and geoRFs, with only temperature (in °C) and wind speed (in m/s) considered. Temporal variation was not considered for this case study.

Mean PM10 recorded for the year 2013 were observed at 40 locations across New Zealand (see Figure 2.2). Significant spatial autocorrelation was identified across the study region. This was confirmed by Moran's I, which was calculated as $I = 0.3577$ with a corresponding p-vale for the two-sided test for presence of spatial autocorrelation of 3.23×10^{-8}.

We computed 45 geoRFs on the PM10 data. The models were determined by varying the neighbourhood structure (fixed, adaptive, or clustered), the bandwidth, and the weight parameter. For the geoRFs that used an adaptive bandwidth to define the neighbourhood, bandwidths of 2, 5, and 8 were tried. The bandwidths used here were relatively small compared to those tried in the simulation study. This reflects that the number of observations in the PM10 data is much smaller. For the geoRFs that used a fixed bandwidth to define the neighbourhood, bandwidths of 15, 60, and 150 km were tried. These bandwidths correspond to the 5th-, 10th-, and 20th-percentiles of all pairwise distances between locations in the PM10 data. Finally, for the geoRFs that used K-means clustering to define the

Table 5.7: Experimental design for fitting geoRF to the New Zealand particulate matter case study.

Kernel	Bandwidth, b	Weight parameter, a	Total
Adaptive	2, 5, 8	0, 0.25, 0.5, 0.75, 1	$3 \times 5 = 15$
Fixed	15, 60, 150	0, 0.25, 0.5, 0.75, 1	$3 \times 5 = 15$
Cluster	2, 3, 4	0, 0.25, 0.5, 0.75, 1	$3 \times 5 = 15$
Total			45 models

neighbourhood, we tried 2, 3, and 4 clusters. The experimental design for the methods are given in Table 5.7.

A 10-fold cross validation was performed. Ten training sets of 32 data observations were randomly sampled from the 40 total, without replacement. For each iteration, the remaining 8 data observations were put aside at the test set. We fitted the 45 models to each of the training sets to train the models, and the test sets were used to compute predictions at different locations and to calculate RMSE, MAE, and Moran's I on the residuals. The mean RMSE, and mean MAE were computed and compared to assess the performance of each model.

Figure 5.9 shows the mean RMSE, mean MAE, and mean Moran's I calculated over each test set for each model. We observed a different trend to what was evidenced in the spatial simulation study. When an adaptive kernel (left three plots in Figure 5.9) was used to select the observations to be included in the neighbourhood of each local RF for each location in the training set, we observed a generally decreasing trend in mean RMSE and mean MAE, as the weight parameter increased from $a = 0$ to $a = 1$. This trend was observed for each bandwidth tried. This suggests that when more observations were included in the local RF for each observation, the more accurate the predictions become. Further, the geoRF produced more accurate predictions on new data than the traditional RF (when $a = 0$), which opposes what we concluded in the simulation study. This may be

due to that fact that when distance between pairs of observations increases in the PM10 data, the less appropriate it is to include them in the prediction process.

In addition to increasing the predictive accuracy for new data, they also accounted for spatial autocorrelation better as the weight parameter increased. A mean Moran's I that is negative and/or far from zero suggests that the model does not properly accounting for spatial autocorrelation. We observed negative mean Moran's I for each model where an adaptive bandwidth was used, for each bandwidth and for each weighting parameter tried. This suggests that spatial autocorrelation may not have been properly accounted for. However, as the weighting parameter increased, we observed a generally increasing trend in mean Moran's I, with the largest observed for the model with a bandwidth of 5 and weight parameter of 1. It appears that when we included more predictions from the local RF on each observation, more spatial autocorrelation was accounted for. We concluded the same trend in the simulation study.

When a fixed kernel was used to select the observations to be included in the neighbourhood of each local RF for each location in the training set, we observed relatively no trend in mean RMSE and mean MAE when the weight parameter was increased, for all bandwidths tried. Once again, this trend was different to what was evidenced in the spatial simulation study. Furthermore, mean RMSE and mean MAE were higher than geoRFs that were computed using adaptive or clustered approaches. This suggests that using a fixed bandwidth in this setting did not increase the predictive accuracy. In addition, mean Moran's I was negative for all fixed bandwidths tried, and did not increase much when the weight parameter increased, suggesting spatial autocorrelation was accounted for poorly.

When a clustering approach was taken to selecting observations to be included in the neighbourhood of each local RF for each cluster in the training set, we observed a similar trend in mean RMSE and mean MAE as the weight parameter increased to that of the adaptive case. We observed

a generally decreasing trend in mean RMSE and mean MAE, when the weight parameter increased. Further, we observed that when the number of clusters was 2, the mean RMSE and mean MAE decreased more so as the weight parameter increased. This suggests that when more observations were included in the local RF for each observation, the more accurate the predictions become.

In addition to increasing the predictive accuracy for new data, the geoRFs using a clustered approach also accounted for spatial autocorrelation better as the weight parameter increased. We observed negative mean Moran's I for each model where clustering was used. This again suggests that spatial autocorrelation may not have been properly accounted for. However, as the weighting parameter increased, we observed a generally increasing trend in mean Moran's I. It appears that when we included more predictions from the local RF on each observation, more spatial autocorrelation was accounted for. We concluded the same trend in the simulation study.

The forest that gave the lowest mean RMSE and mean MAE was the one that used a clustered neighbourhood structure, with $K = 2$, and had a weight parameter of $a = 1$. Mean RMSE and MAE were calculated to be 2.60 and 2.05, respectively, and mean Moran's I was calculated to be $I = -0.0863$. Moran's I calculated on the residuals indicated that this model accounted for spatial autocorrelation reasonably well. We choose this model as the best to compute predictions for new data.

In order to construct an interpolated surface map of PM10, we will use temperature and wind speed observations from 347 locations across New Zealand in 2013 to calculate predicted values from a geoRF. A map of the locations is shown in Figure 5.10. To compute the predicted values, we refit the geoRF using all 40 data observations, with a clustered neighbourhood where $K = 2$. We used a weighting parameter of $a = 1$, which meant predictions were calculated using only the local sub models of observations within their corresponding cluster. Figure 5.11 displays the interolated surface map of PM10 across New Zealand, using the geoRF with

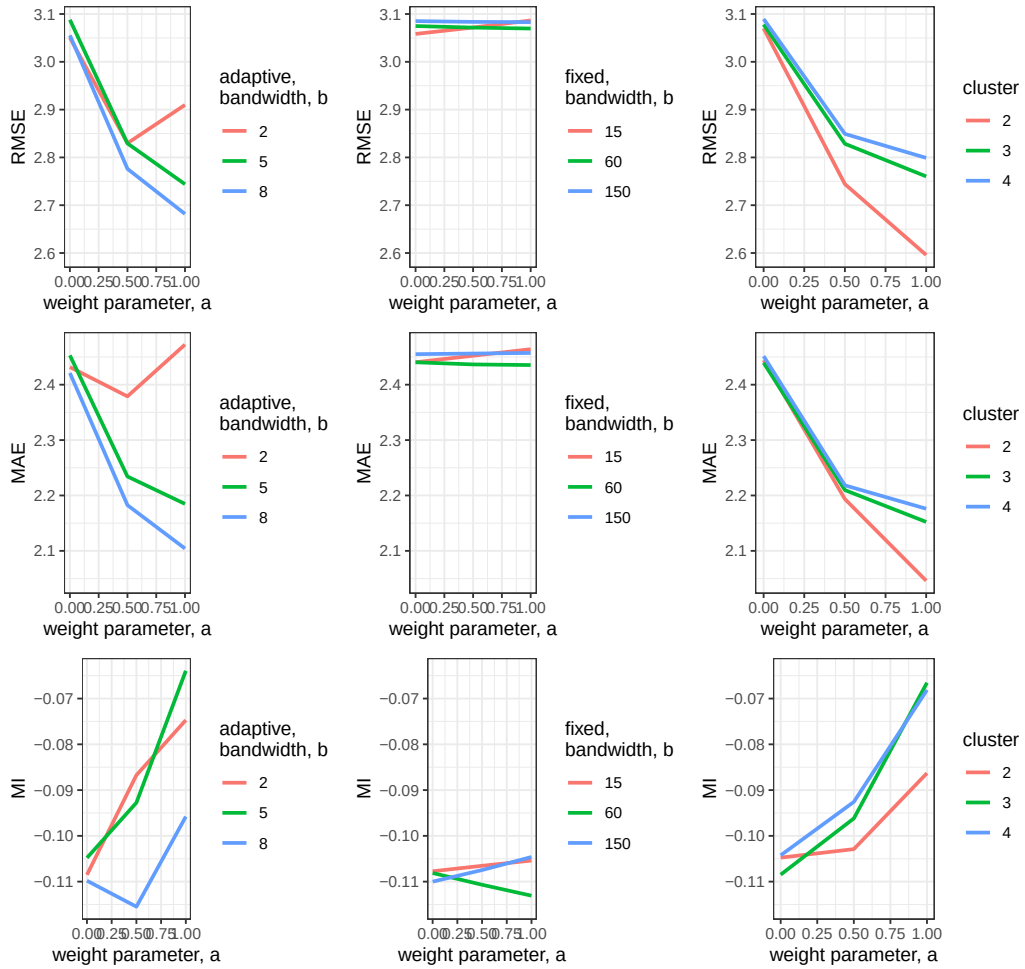


Figure 5.9: Mean measures of accuracy and spatial autocorrelation for each approach. On the left, an adaptive kernel approach was used for the geoRF, with bandwidths 3, 5, and 8 tried. In the middle, a fixed kernel approach was used for the geoRF, with bandwidths 15, 60, and 150 km tried. On the right, our clustering approach was used for the geoRF, with 2, 3, and 4 clusters tried. The weighting parameter is displayed on the x-axis.

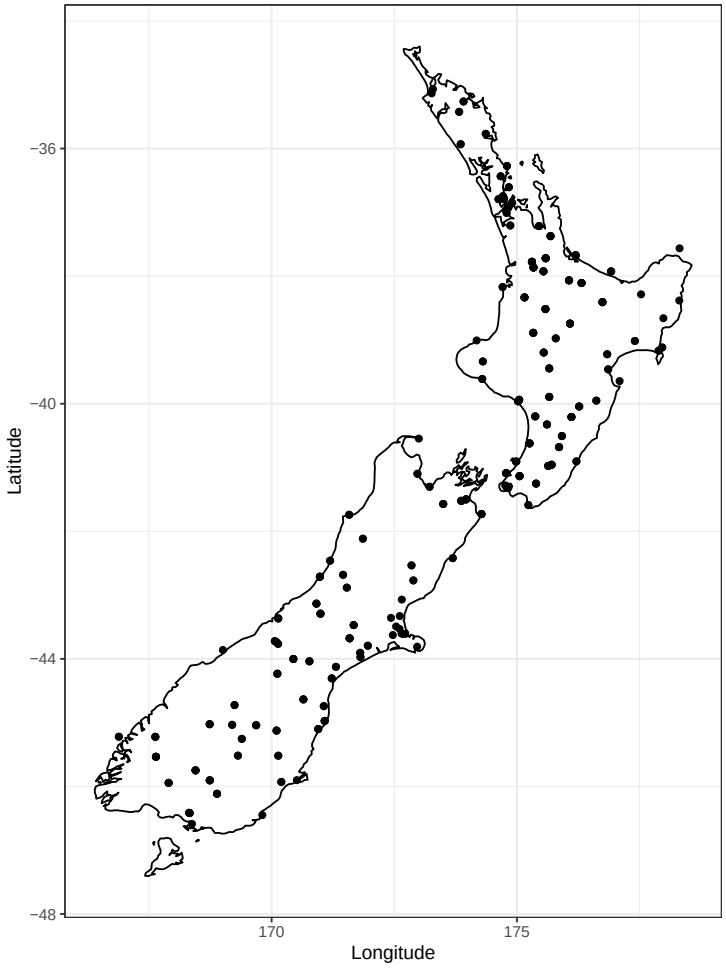


Figure 5.10: Locations of the stations that recorded temperature and wind speed across New Zealand.

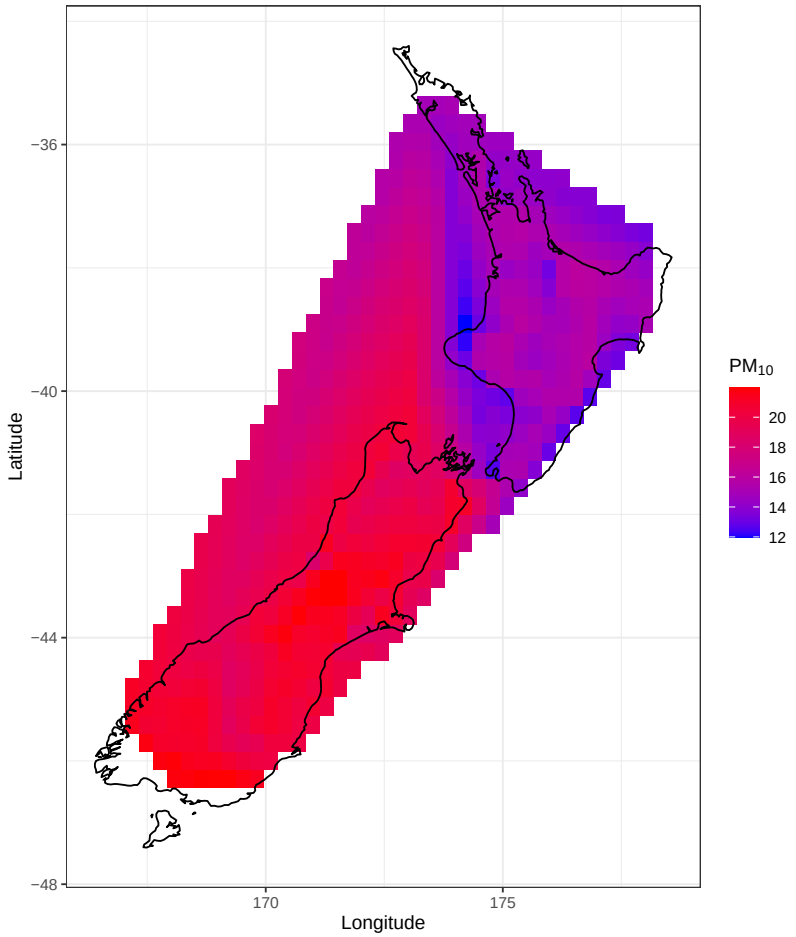


Figure 5.11: Surface plot for the predicted annual PM10 concentration in New Zealand for 2013.

a $K = 2$ clustered neighbourhood and weight parameter of $a = 1$. We observe higher concentrations of PM10 in the southern part of the country, compared to lower values in the northern part. The case study validates the usefulness of the clustered geoRF technique for spatial data.

Table 5.8: Number of observations, n_t of hoki catch weight and the training (u_t) and test (v_t) set sizes, for each year. The ratio of training data to test data was approximately 4:1.

Year	n_t	u_t	v_t
2000	100	80	20
2001	100	80	20
2002	98	78	20
2003	77	61	16
2004	85	57	17
2005	88	70	18
2006	88	70	18
2007	88	70	18
2008	90	72	18

5.7.2 Sub-Antarctic hoki

We performed a case study, in which we computed RF and geoRF on the sub-Antarctic hoki catch weight data described in Section 2.7.2. The aim was to find the best forest to predict hoki catch weight across the sub-Antarctic region, in terms of predictive accuracy RMSE, and MAE, while also accounting for temporal and spatial autocorrelation. Limited covariates were available to compute the RF and geoRFs, with only depth (in m), stratum, and year considered.

Observed hoki catch weight in kilograms was recorded for 814 trawls taken throughout the sub-Antarctic region, for the years 2000 – 2008 (see Figure 1.4). The number of observations within each year changed, and the locations of the trawls were different each year. Table 5.8 gives the number of observations within each year of the hoki data.

Significant spatial autocorrelation was identified across the study region for most years, as evidenced by the interpolated surface plots in Figure 2.2. This was confirmed by Moran's I, which was calculated for each year.

Table 5.9: Moran's I and p-values for the two-sided test for presence of spatial autocorrelation for hoki catch weight observed over the sub-Antarctic region for the years 2000 – 2008.

Year	I	P-value
2000	0.0727	1.55×10^{-6}
2001	0.1866	0.143
2002	-6.078×10^{-4}	0.615
2003	0.1932	$< 2.2 \times 10^{-16}$
2004	0.0681	5.19×10^{-9}
2005	0.0946	3.59×10^{-5}
2006	0.1132	6.85×10^{-8}
2007	-4.561×10^{-3}	0.537
2008	5.790×10^{-3}	0.442

Table 5.9 gives the Moran's I and corresponding p-values for the two-sided test for presence of spatial autocorrelation.

We computed geoRFs on the hoki data as a proof of concept for the geoRF methodology applied to spatio-temporal data. Unlike the PM10 case study or the simulation studies, models were determined by varying the neighbourhood structure (fixed, adaptive, or clustered), and the weight parameter only. The bandwidth and temporal correlation parameters were specified *a priori* to focus on effect that changing the neighbourhood and weight parameter has on predictive accuracy, only. For the geoRFs that used an adaptive bandwidth to define the neighbourhood, a bandwidths of 50 was used. This was thought to ensure that a reasonable number of observations be included in the local sub models. For the geoRFs that used a fixed bandwidth to define the neighbourhood, a bandwidths of 200 km was used. This corresponded to the 25th-percentile of all pairwise distances between locations in the hoki data. Finally, for the geoRFs that used K-means clustering to define the neighbourhood, we used 10 clusters. The

Table 5.10: Experimental design for fitting geoRF to the sub-Antarctic hoki case study.

Kernel	Bandwidth, b	Weight parameter, a	Temporal correlation, ρ	Total
Adaptive	50	0, 0.5, 1	0.7	$1 \times 3 \times 1 = 3$
Fixed	200	0, 0.5, 1	0.7	$1 \times 3 \times 1 = 3$
Cluster	10	0, 0.5, 1	NA	$1 \times 3 = 3$
Total				9 models

experimental design for the methods are given in Table 5.10.

A 10-fold cross validation was performed. Training sets were randomly sampled from within the set of hoki data for each year, without replacement. For each year, the remaining data observations were put aside at the test set. This was repeated ten times. We fitted the 9 models to each of the training sets to train the models, and the test sets were used to compute predictions at different locations and to calculate RMSE, MAE, and Moran's I on the residuals. The mean RMSE, and mean MAE were computed and compared to assess the performance of each model.

Figure 5.12 displays the mean RMSE and mean MAE calculated over each test set, for each model. When an adaptive kernel was used to select the observations to be included in the neighbourhood of each local RF for each observation in the training set, we observed that the geoRF resulting in the lowest mean RMSE and mean MAE was that when a weighting parameter of $a = 0.5$ was used. This suggests that when we compute the predictions weighting the global and local random forests equally, we improve the predictive accuracy on new data. The same trend was observed for the fixed kernel approach. When a clustered approach was taken, the traditional RF methodology (when $a = 0$ resulted in the lowest mean RMSE and mean MAE.

Figure 5.13 displays the mean Moran's I calculated from the residuals over each test set, for each model, and for each year. We observe erratic trends

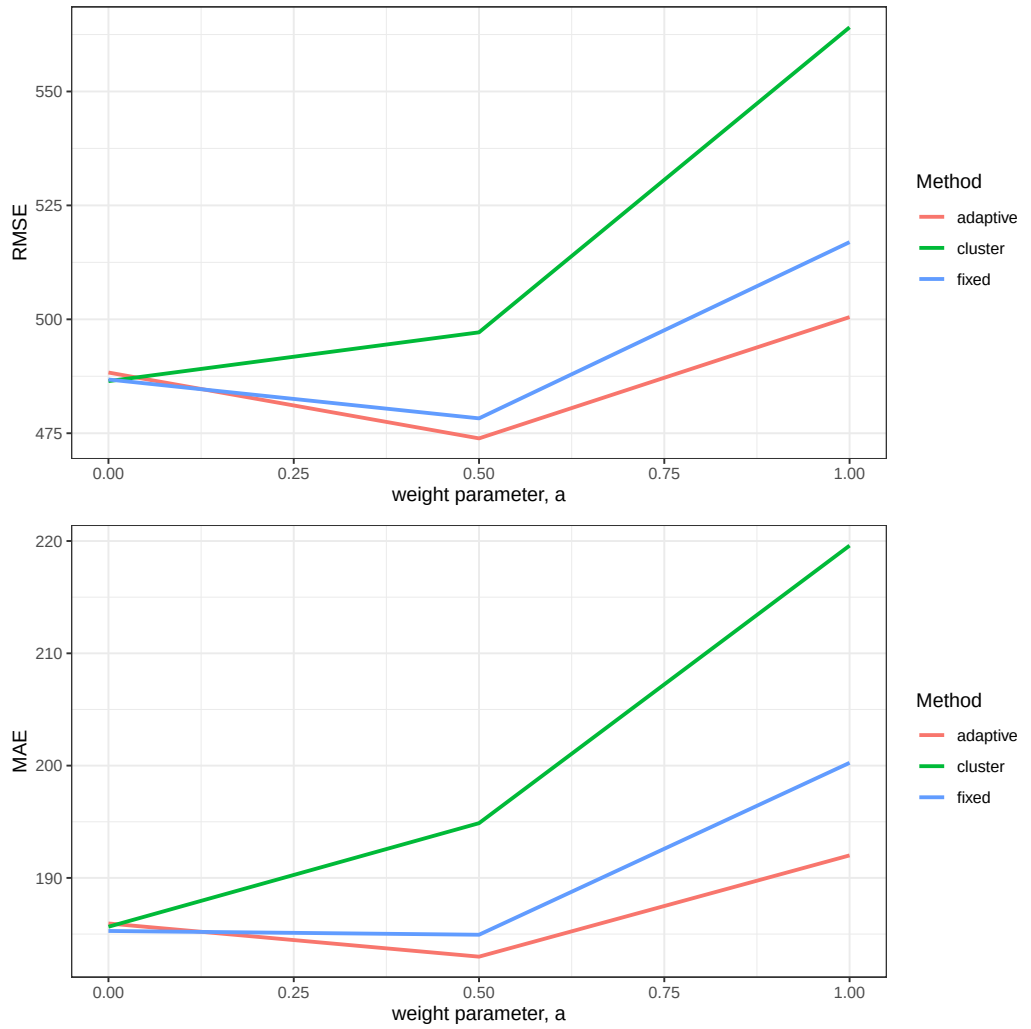


Figure 5.12: Mean measures of accuracy for each approach. The weighting parameter is displayed on the x-axis.

in the mean Moran's I across both kernels, and weighting parameters. We conclude that spatial autocorrelation was accounted for in some years, and in others, was not accounted for properly. We hypothesize that this is due to the misspecification of temporal autocorrelation in the geoRF, and might be solved by exploring different correlation parameters.

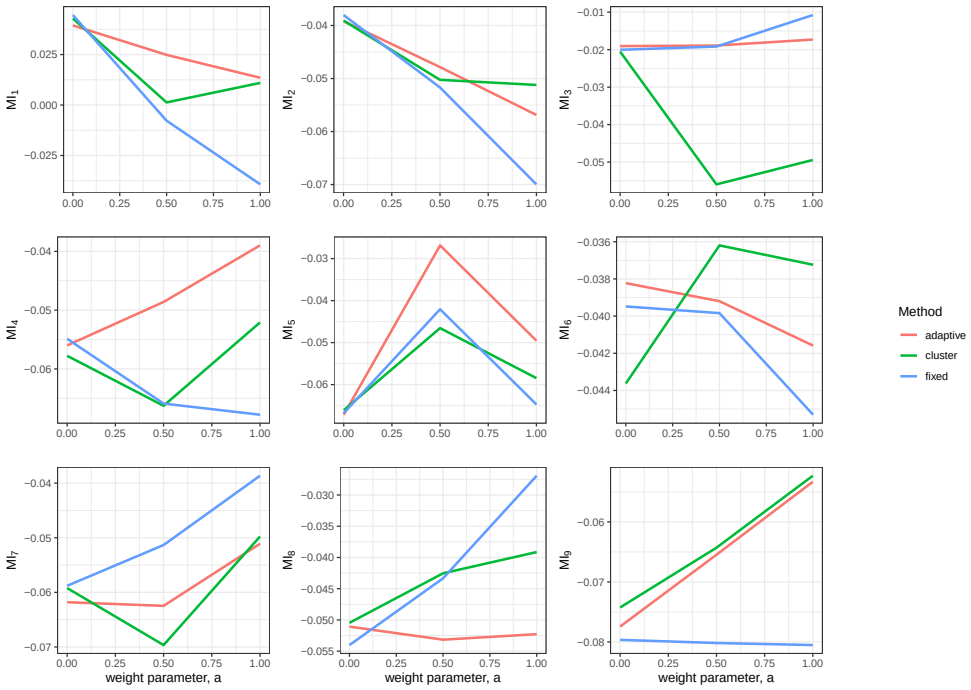


Figure 5.13: Mean measures of spatial autocorrelation for each approach, for each year. The weighting parameter is displayed on the x-axis.

5.8 Conclusion

In this chapter, we proposed an extension of the geographic random forest methodology to incorporate neighbourhood structures that were constructed via K-means clustering. Further, we extended the methodology to spatio-temporal point referenced data. The geographic random forest method has the benefit over parametric modelling approaches because of the lack of needing to provide a covariance structure. However, we did not find improvements in predictive accuracy when K-means clustering was used to define the neighbourhood structure.

Chapter 6

Pivotal discrepancy measures for Bayesian modelling of spatio-temporal data

As we have seen throughout this thesis, the literature that surrounds the subject of spatial and spatio-temporal statistics is predominantly concerned with parametric inference for the covariance structure. Within geostatistics (where data are observed at specific locations in time), a great deal of attention has been paid to proposing and describing new spatial and spatio-temporal models, studying their characteristics, and developing estimation methods from within both frequentist and Bayesian frameworks. Markedly less attention has been paid to developing goodness-of-fit tests that identify model misspecification or allow for selection of a "best" model. Model misspecification in the context of parametric covariance models for spatio-temporal processes means models that have an incorrect mean function or covariance structure. At present, there is no generalized formal theory for assessing goodness-of-fit for spatio-temporal models that are defined using parametric covariance functions. Instead, there is a range of criteria and tests that have been used when fitting a spatio-temporal covariance model to data.

Literature has seen the use of Akaike information criterion, AIC (Akaike (1973)), and Bayesian information criterion, BIC (Schwarz et al. (1978)), which are popular model selection tools for a wide range of frequentist and Bayesian statistical applications and models. Huang et al. (2007) proposed model comparison for space-time models using these criteria, and investigated their usefulness through simulation and an application to surface shortwave radiation budget analysis. Another criterion, deviance information criterion, DIC (Spiegelhalter et al. (2002)), is used by Pollice (2011) to compare multivariate receptor models for identifying the spatial locations of major PM10 pollution sources. To compare predictive capabilities, mean squared and root mean squared prediction errors at fixed times can be calculated and this is illustrated in Huang et al. (2007). Further, Sahu & Bakar (2012) applied the predictive model choice criterion (PMCC), which included a term for model complexity. In more recent times, we have seen the proposal and use of widely applicable information criterion, WAIC, Watanabe (2010), an information criterion constructed in the same vain as DIC, but fully Bayesian. Vehtari & Gelman (2014) and Vehtari et al. (2017) adopted WAIC as a method for approximating leave-one-out cross validation for model goodness-of-fit.

These model selection/goodness-of-fit criterion are inappropriate for some spatio-temporal models. AIC, BIC, DIC, and WAIC all require that the joint data likelihood be calculated as the product of the marginal data likelihoods. However, this assumption breaks the spatial and temporal dependence structure at the lower levels of the hierarchical model.

A promising methodology for assessing goodness-of-fit for Gaussian random fields (GRFs) in the Bayesian framework is that of pivotal discrepancy measures, introduced in Johnson (2007) and further investigated in Jun et al. (2014). The approach can be used for GRFs with stationary and nonstationary covariances and to data observed at regular or irregularly spaced locations. In this paper, we extend the approach described in Jun et al. (2014) for assessing goodness-of-fit of Bayesian spatio-temporal mod-

els using pivotal discrepancy measures. Jun et al. (2014) proposed the method of partitioning the data to increase the power to detect model misspecification, assuming equal partition sizes. In this chapter we make a contribution and extend the Jun et al. (2014) approach to partitions of unequal sizes and the use of K-means partitioning as a method for inducing homogeneity within partitions, when there are no preset spatial boundaries.

Chapter 6 is divided in to the following sections. We first present the pivotal discrepancy measure for a spatio-temporal model evaluated at a sample from the posterior parameter distribution. Further, we present the pivotal discrepancy measure for subset data of unequal size. This is followed by a section that is dedicated to investigation of the usefulness of the test using a simulation study. Following, is an application and evaluation of spatio-temporal models to hoki catch weight data.

6.1 Pivotal discrepancy measure

Assume that a spatio-temporal geostatistical model (Equation 2.24) is fitted to $\boldsymbol{y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_T\}$. Then, let $\tilde{\boldsymbol{\theta}}^{(l)}$ represent the lth draw of the parameter vector $\boldsymbol{\theta}$ from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. We can construct a pivotal quantity,

$$S(\boldsymbol{y}_t, \tilde{\boldsymbol{\theta}}^{(l)}) = (\boldsymbol{y}_t - \boldsymbol{\mu}_t^{(l)})' \Big(\frac{\sigma^{2(l)}}{1 - \rho^{(l)2}} \boldsymbol{R}^{(l)} + \tau^{2(l)} \boldsymbol{I}_n \Big)^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_t^{(l)}) \sim \chi^2_n, \quad (6.1)$$

for $t = 1, ..., T$, $l = 1, ..., L$, and L is the total number of posterior draws. Then $S(\boldsymbol{y}_t, \tilde{\boldsymbol{\theta}}^{(l)})$ is χ^2-distributed on n degrees of freedom Johnson (2007). Jun et al. (2014) highlighted two complications that arise when the spatial equivalent of Equation 6.1 is used in a Bayesian goodness-of-fit test for spatial models, which also arise for the spatio-temporal case. The first complication is that a test based on the test statistic in Equation 6.1 typically provides little power to detect model misspecification when it is applied globally to the entire data vector. Jun et al. (2014) illustrated this

through an example where a simple Bayesian linear regression is fitted to a fictional dataset that exhibited larger variability at the extreme values of a covariate, and smaller variability around the mean of the covariate. Their test based on the spatial equivalent of the test statistic in Equation 6.1 was unable to detect departure of the model from the data. This was due to the cancellation of the large and small contributions from the residuals when the statistic $S(\boldsymbol{y}_t, \tilde{\boldsymbol{\theta}}^{(l)})$ was applied to the entire data set. Jun et al. (2014) proposed a partitioning strategy, where the chi-squared diagnostic was constructed using residuals from distinct regions of the spatial domain. Use of the partitioning strategy allowed the lack of fit of the model to the data in each partition to be correctly detected and overall, the goodness-of-fit test failed. Partitioning of the data was further motivated in a simulation test and applications to Colorado precipitation data and total column ozone data. We propose an extension of their strategy in Section 6.1.1.

The second complication is how to combine the pivotal discrepancy measures based on many posterior draws and a partitioned dataset when conducting a goodness-of-fit test. A single posterior draw, $\tilde{\boldsymbol{\theta}}^{(l)}$, from the posterior distribution based on a non-partitioned dataset gives the statistic $S(\boldsymbol{y}_t, \tilde{\boldsymbol{\theta}}^{(l)}) \sim \chi^2_n$. Each posterior draw gives a different value of the test statistic and these values will be correlated. Jun et al. (2014) proposed diagnostics based on bounds on the distribution of order statistics proposed by Caraux & Gascuel (1992); Rychlik (1992) to carry out a goodness-of-fit test that makes use of the multiple correlated statistics obtained from the posterior draws. We adopt that approach in this article.

6.1.1 Partitioning the observed locations into K subsets (not necessarily of equal size)

Jun et al. (2014) proposed partitioning the set of observed locations into K subsets of size w and showed that partitioning the observation vector into

regions of high and low variability allowed the test to detect model misspecification. They suggest partitioning based on either prior knowledge regarding regions of likely homogeneity, or according to well defined spatial boundaries. Applying Equation 6.1 to the partitioned spatio-temporal data gives:

$$S_j(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)}) = (\boldsymbol{y}_{tj} - \boldsymbol{\mu}_{tj}^{(l)})' \Big(\frac{\sigma^{2(l)}}{1-\rho^{(l)2}} \boldsymbol{R}_j^{(l)} + \tau^{2(l)} \boldsymbol{I}_{w_j} \Big)^{-1} (\boldsymbol{y}_{tj} - \boldsymbol{\mu}_{tj}^{(l)}) \sim \chi^2_w, \quad (6.2)$$

for $t = 1, ..., T$, $j = 1, ..., K$, and $l = 1, ..., L$, where $\boldsymbol{y}_{tj}$, $\boldsymbol{\mu}_{tj}$, and $\boldsymbol{R}_j$ denote the parts of Equation 2.29 corresponding to subset j, and w is the number of observed locations in each subset. When the subsets vary in size we no longer have identical distributions for the pivotal statistics and $S_j(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)}) \sim \chi^2_{w_j}$, where w_j is the number of samples in subset j.
We agree that partitioning is necessary to improve the performance of the goodness-of-fit test. However, we allow the subsets to vary in size and we adopt the K-means clustering algorithm (Algorithm 3) to partition the spatial domain into regions of likely homogeneity.

6.1.2 Nominal distribution of the ordered pivotal statistics

The screening diagnostics we use are based on bounds of order statistics given in Proposition 3 in Caraux & Gascuel (1992). These bounds are applied to non-identically distributed dependent variables and we thus generalise the diagnostics proposed by Jun et al. (2014).
Let $X_{(1)}, ..., X_{(N)}$ denote a set of order statistics from a dependent sample of N random variables with non-identical distribution functions, $F_{X_1}, ..., F_{X_N}$. Also, let $F_{x_{r:N}}$ denote the distribution function for the rth-order statistic out of a sample of N dependent draws from $F_{X_1}, ..., F_{X_N}$. Then,

$$\sup\left(0, 1 - \frac{\sum_{j=1}^{N}\big(1 - F_{X_j}(x)\big)}{N-r+1}\right) \leq F_{X_{r:N}}(x) \leq \inf\left(\frac{\sum_{j=1}^{N} F_{X_j}(x)}{r}, 1\right). \quad (6.3)$$

We partition the spatial domain into K groups of potentially unequal size, $w_j, j = 1, \ldots, K$. The pivotal statistic $S_j(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)})$ is calculated for each partition, $j = 1, \ldots, L$, time point $t = 1, \ldots, T$ and posterior draw $l = 1, \ldots, L$.

This results in a total of KTL dependent test statistics $\{S_j(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)}) : j = 1, ..., K; t = 1, ..., T, l = 1, ..., L\}$, with respective density function $\chi^2_{w_j}$. We denote the rth-order statistic from this set by $S_{(r)}$, where $r = 1, ..., KTL$, and let F_r denote the distribution function of the $\chi^2_{w_j}$ distribution. It follows from above that,

$$P(S_{(r)} < t) \leq \inf\left(1, \frac{\sum_{r=1}^{KTL} F_r(t)}{r}\right),$$
$$P(S_{(r)} > t) \leq \inf\left(1, \frac{\sum_{r=1}^{KTL} \big(1 - F_r(t)\big)}{KTL - r + 1}\right).$$

6.1.3 Pivotal discrepancy measure goodness-of-fit test for Bayesian inference

We propose the following procedure for testing goodness-of-fit for Gaussian spatio-temporal models:

1. Partition the set of observed locations into K subsets, Q_j of size w_j, where $j = 1, ..., K$, using K-means clustering. For each $t = 1, ..., T$, let $\boldsymbol{y}_{tj}$, $\boldsymbol{X}_{tj}$, and $\boldsymbol{R}_j$ denote the parts of Equation 2.29 that correspond to subset Q_j.

2. Generate posterior samples for $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(L)}$, based on the complete observed data $(\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$.

3. For every sampled parameter vector $\boldsymbol{\theta}^{(l)}$, and each data subset $\boldsymbol{y}_{tj}$, for every t, calculate the pivotal statistic in Equation 6.2.

4. Collect all KTL statistics in an ordered set $\{S_j(\boldsymbol{y}_{tj}, \boldsymbol{\theta}^{(l)}) : j = 1, ..., K, t = 1, ..., T, l = 1, ..., L\}$, and denote the rth-order statistic from this set by $S^*_{(r)}$.

5. Perform the two-sided goodness-of-fit test of significance level α by specifying integers m and u such that $1 \leq m < u \leq KTL$, and deter-

mining t_m and t_u such that,

$$\left(\left[\frac{\sum_{k=1}^{KTL} F_k(t_m)}{m}\right] - \frac{\alpha}{2}\right)^2, \tag{6.4}$$

and

$$\left(\left[1 - \frac{\sum_{k=1}^{KTL}\left(1 - F_k(t_u)\right)}{KTL - u + 1}\right] - \frac{\alpha}{2}\right)^2, \tag{6.5}$$

are minimized. If either $S^*_{(m)} < t_m$ or $S^*_{(u)} > t_u$, then the assumed model can be rejected in a two-sided test of size α.

Jun et al. (2014) recommend that m and u be selected such that $m = r_m KTL$ and $u = r_u KTL$, where $0 < r_m < r_u < 1$, for example $r_m = 0.1$ and $r_u = 0.9$.

6.2 Simulation

A simulation experiment was performed to assess the ability of the goodness-of-fit test to detect misspecification of the covariance structure of a model. A total of 30 pairs of longitude and latitude values, $\boldsymbol{s} = (s_1, s_2)$, were sampled randomly from one of three subsets within the unit square. Within the first subset, S_1, five locations were generated uniformly from the lower left [0, 0.2] $\times$ [0, 0.2] portion of the unit square. In the second subset, S_2, ten locations were uniformly sampled from the lower right [0.8, 1] $\times$ [0, 0.2] portion of the unit square. Finally, in subset S_3, fifteen locations were uniformly sampled from the entire unit square. The motivation is that the fit of a covariance model can be best tested by comparing its fit in distinct regions (where its local smoothness properties can be evaluated), with its fit to point distributed throughout the domain (where its global features can be evaluated) as mentioned in Jun et al. (2014). Subsets S_1 and S_2 provided clusters of locations that allow for the assessment of local model fit, whereas subset S_3 provides motivation for assessing global model fit. Figure 6.1 shows the simulated locations and the corresponding subsets.

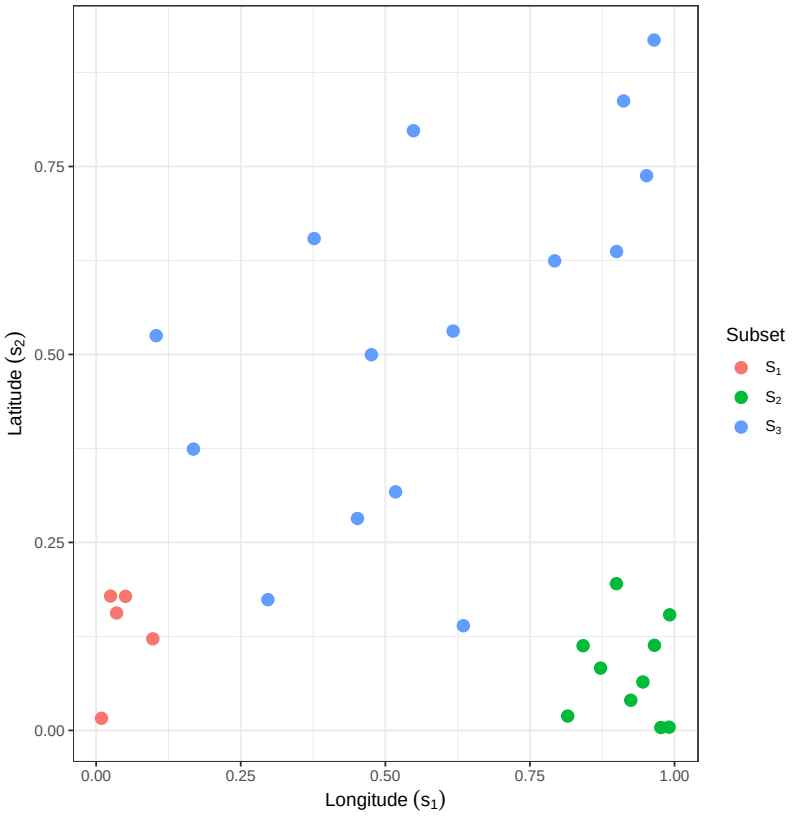


Figure 6.1: Domain and locations of the simulated data, colour-coded by subset

Three datasets were simulated using the spatio-temporal process defined in Section 2.4.2. The mean process, $\boldsymbol{\mu}_t$, was set to zero, to allow for detection of model misspecification through the covariance structure only. Observed data $\{\boldsymbol{y}_t\}$ were simulated for $t = 1, ..., 5$ time units.

Three variants of the Matèrn correlation function with closed form expressions were used to construct the covariance matrix $\sigma^2\boldsymbol{R}$. The first variant,

$$(\boldsymbol{R})_{ij} = \exp\left(\frac{-||\boldsymbol{s}_i - \boldsymbol{s_j}||}{\phi}\right), \tag{6.6}$$

is the closed form of the Matèrn correlation function, where the smoothness parameter, ν, is set to 0.5 and is also known as the exponential correlation function. The second variant,

$$(\boldsymbol{R})_{ij} = \exp\left[-\left(\frac{||\boldsymbol{s}_i - \boldsymbol{s_j}||}{\psi}\right)^2\right], \tag{6.7}$$

is the closed form of the Matèrn correlation function, where the smoothness parameter, $\nu \to \infty$, and is known as the Gaussian correlation function. The third variant,

$$(\boldsymbol{R})_{ij} = s_{2i}s_{2j}\exp\left(\frac{-||\boldsymbol{s}_i - \boldsymbol{s_j}||}{\phi}\right), \tag{6.8}$$

is a non-stationary form of the exponential correlation function given by Equation 6.6, that allows the correlation between observations separated by a distance d to scale by their latitudes, s_2.

The following parameters were chosen to simulate the data $\{\boldsymbol{y}_t\}$. The measurement variance (nugget variance), $\tau^2 = 0.0001$, and the spatio-temporal variance $\sigma^2 = 1$. We chose $\sigma^2 > \tau^2$ to focus on identifying incorrect spatio-temporal covariance structure. Further, we set $\rho = 0.7$ to induce a moderately positive temporal autocorrelation that might be observed in reality. Finally, we set $\phi = 0.2$ in Equations 6.6 and 6.8, and $\psi = 0.8$ in Equation 6.7 to induce spatial autocorrelation. The specific values were chosen because they produced data that exhibited spatial autocorrelation.

We fitted the spatio-temporal geostatistical model given in Section 2.4.2 with the covariance function in Equation 6.6 to each of the three datasets. We excluded covariates, with only a single intercept term included in the mean function, such that $\boldsymbol{\mu}_t = \mathbf{1}_{30}\beta$, where 1_{30} is a vector of 1's. The parameters $\boldsymbol{\theta} = (\beta, \tau^2, \sigma^2, \phi, \rho)'$ were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta \sim \text{N}(0, 100),\ \tau^2 \sim \text{IG}(2, 1),\ \sigma^2 \sim \text{IG}(2, 1),\ \phi \sim \text{U}(0.001, 2),\ \rho \sim \text{U}(-1, 1).$$

Markov chain Monte Carlo (MCMC) was used to fit the model to the data and this was done through **R** using the package `NIMBLE` (NIMBLE Development Team (2017)). Two chains, each 100000 iterations, were generated of the parameter vector $\boldsymbol{\theta} = (\beta, \rho, \phi, \sigma^2, \tau^2)'$ for each dataset. The first 90000 iterations from each chain were discarded as warm-up, and the remaining draws were combined, resulting in a posterior sample of size $L = 20000$. For each fitted model, pivotal quantities for every posterior sample were calculated inline with equation 6.2. We considered three cases of partitioning to assess the impact it has on testing goodness-of-fit. In the first case, the locations were not partitioned into subsets. Pivotal quantities for each fitted model $S(\boldsymbol{y}_t, \tilde{\boldsymbol{\theta}}^{(l)})$ for $t = 1, ..., 5$ and $l = 1, ..., 20000$ were calculated, combined and ordered. For the second case, the locations were partitioned into $K = 3$ subsets of $w = 10$. Pivotal quantities for each fitted model $S(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)})$ for $t = 1, ..., 5$, $j = 1, 2, 3$, and $l = 1, ..., 20000$ were calculated, combined and ordered. Finally, the locations were partitioned into the subset S_1, S_2, and S_3, that were used to simulate the locations. Pivotal quantities for each fitted model $S(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)})$ for $t = 1, ..., 5$, $j = 1, 2, 3$, and $l = 1, ..., 20000$ were calculated, combined and ordered.

Table 6.2 gives the 10th and 90th percentiles of the aggregated (over time, and subset) ordered pivotal discrepancy measures for the model applied to each of the three data sets, in each of the three cases of subsetting. In order to confirm that the model provided a good fit, the quantities need to be within the interval given by the critical values that are calculated from the nominal χ^2 distributions assumed when they were calculated.

Table 6.1: Summary statistics for the model fitted to each simulated data set.

			Dataset 1
	True	Median	(95% CI)
β	0	-0.01656	(-0.2026, 0.1802)
ϕ	0.2	0.2025	(0.1108, 0.3555)
ρ	0.7	0.8370	(0.7451, 0.9115)
σ^2	1	0.6512	(0.4125, 1.020)
τ^2	0.0001	0.1362	(0.07338, 0.2186)
			Dataset 2
β	0	-0.006492	(-0.2075, 0.1981)
ϕ	0.8	1.857	(1.517, 2.000)
ρ	0.7	0.6470	(0.5042, 0.7795)
σ^2	1	0.4947	(0.3449, 0.6768)
τ^2	0.0001	0.03150	(0.02274, 0.04160)
			Dataset 3
β	0	-0.01282	(-0.1984, 0.1651)
ϕ	0.2	0.3853	(0.1542, 0.8944)
ρ	0.7	0.6048	(0.3959, 0.7860)
σ^2	1	0.2532	(0.1219, 0.4708)
τ^2	0.0001	0.06681	(0.04268, 0.09825)

Table 6.2: The 10th and 90th percentiles of ordered pivotal discrepancy measures for the model applied to each of the three datasets. These are compared to the nominal 10th and 90th percentiles: 12.76 and 56.33 for the non-subset data; 1.827 and 27.11 for the even subset data; 0.4894 and 31.71 for the uneven subset data.

	Nominal percentiles	Dataset 1	Dataset 2	Dataset 3
No subset	(12.76, 56.33)	(18.77, 42.21)	(15.33, 27.77)	(15.65, 35.04)
Even subset	(1.827, 27.11)	(3.945, 22.71)	(5.222, 40.04)	(0.4609, 12.54)
Uneven subset	(0.4894, 31.71)	(0.9661, 31.50)	(2.038, 60.47)	(0.1575, 19.65)

In the first case, when no the locations were not partitioned in order to calculate the pivotal quantities, it was found that the 10th and 90th percentiles of ordered pivotal discrepancy quantities were within the corresponding nominal percentiles of (12.76, 56.33). This suggests that the model provided a good fit to each of the three simulated data sets. In the second case, when the locations were partitioned into 3 even subsets to calculate the pivotal quantities, it was found that the 10th and 90th percentiles of the aggregated ordered pivotal discrepancy quantities were within the corresponding nominal percentiles of (1.827 and 27.11) when the model was applied to data set 1. However, the 10th and 90th percentiles of the aggregated ordered pivotal discrepancy quantities were outside the nominal percentiles when the model was applied to data sets 2 and 3. This suggests that the model provides a good fit only to data set 1. A similar result was observed for the final case, with the locations being partitioned into three uneven subsets. It was found that the 10th and 90th percentiles of the aggregated ordered pivotal discrepancy quantities were within the corresponding nominal percentiles of (0.4894 and 31.71) when the model was applied to data set 1. However, the 10th and 90th percentiles of the aggregated ordered pivotal discrepancy quantities were outside the nominal percentiles when the model was applied to data sets

2 and 3. This suggests that the model provides a good fit only to data set 1.

We would have expected that the model only provide a good fit to data set 1, because the model used to generate that data set is the same as the one being fitted. This is correctly executed in the two cases of partitioning. In the first case, the model provided a good fit to each data set, because a lack of partitioning caused a decrease in power to detect the differences. This is highlighted in Figures 6.2 – 6.4. In those Figures, the pivotal discrepancy quantities from each model applied to each data set in each case of partitioning are plotted as a density, and are overlaid with the nominal densities. We see for each data set that when no partitioning occurs, there is sufficient overlap of the pivotal quantities observed and the nominal densities to suggest the model provides a good fit. This is also the case for the partitioning scenarios for data set 1, but not the case for data sets 2 and 3.

6.3 Case study

6.3.1 Hoki catch data from sub-Antarctic survey

We performed a case study, in which we fitted several models to the gridded sub-Antarctic hoki catch weight data described in Section 2.7.2. The aim was to use partitioned and pivotal discrepancy measures to assess goodness-of-fit of each model.

Observed hoki catch weight in kilograms was recorded for 814 trawls taken throughout the sub-Antarctic region, for the years 2000 – 2008 (see Figure 1.4). The number of observations within each year changed, and the locations of the trawls were different each year. However, the CRN models that we developed in this thesis for point reference spatio-temporal data can only be applied to data that were observed at the same locations throughout time. Due to this fact, we gridded the hoki data according to

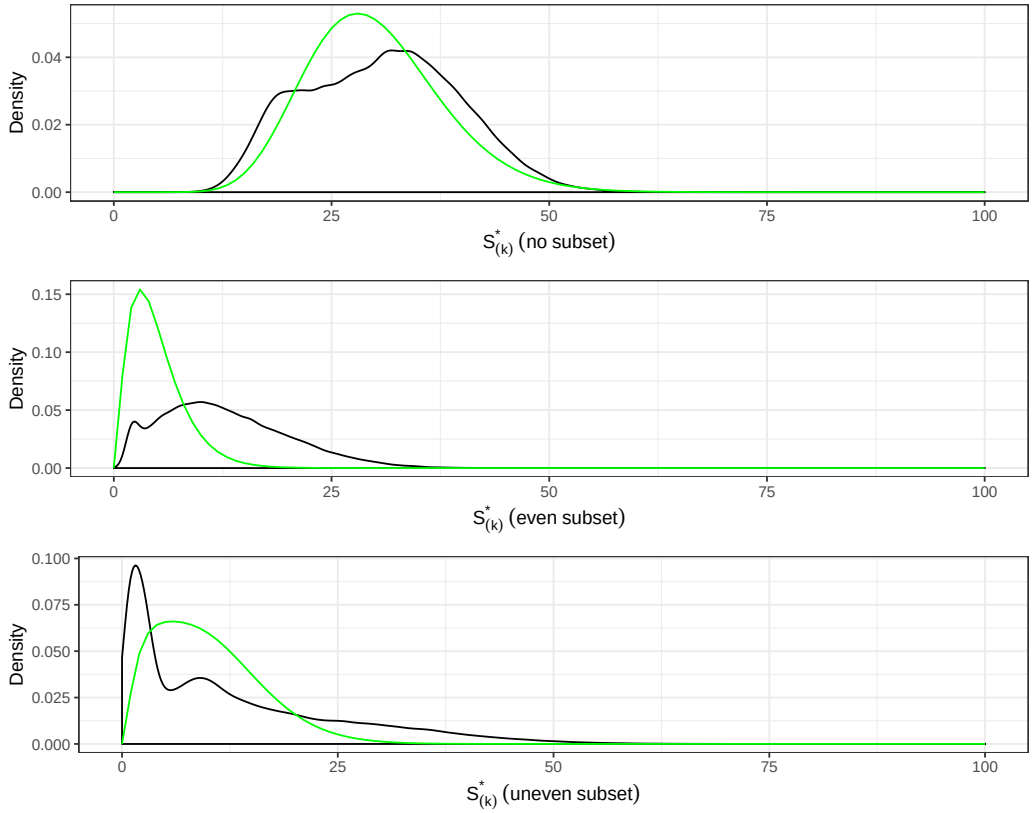


Figure 6.2: PDM density for data set 1

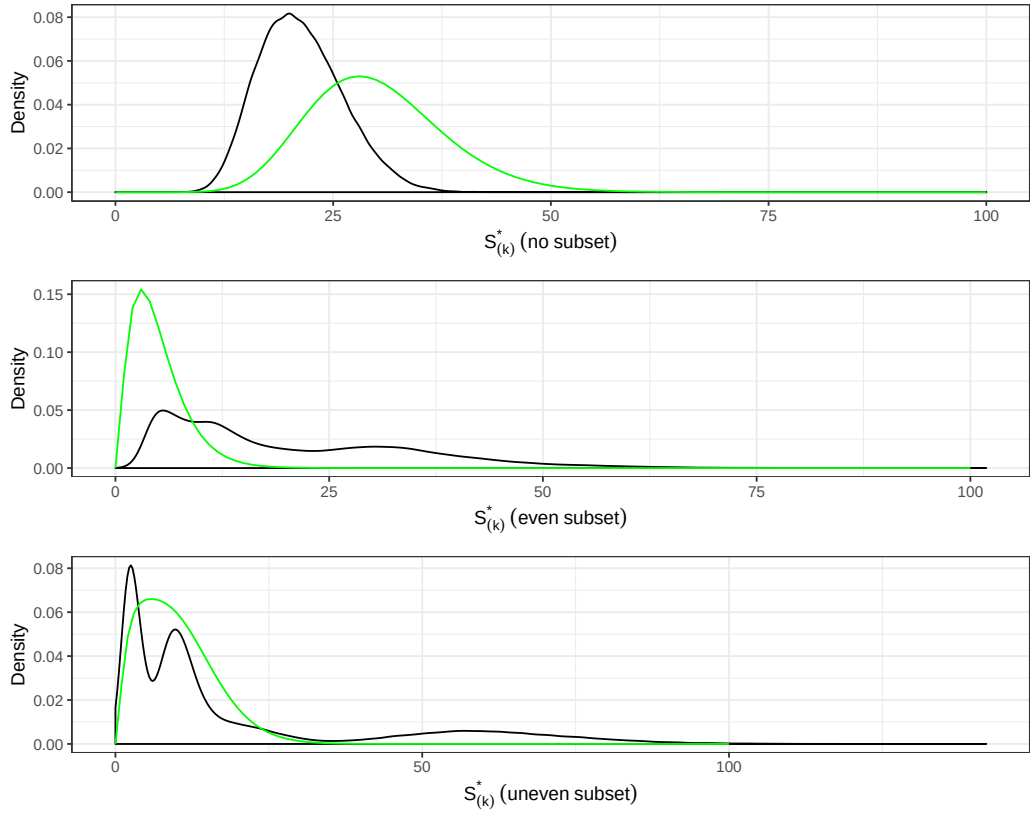


Figure 6.3: PDM density for data set 2

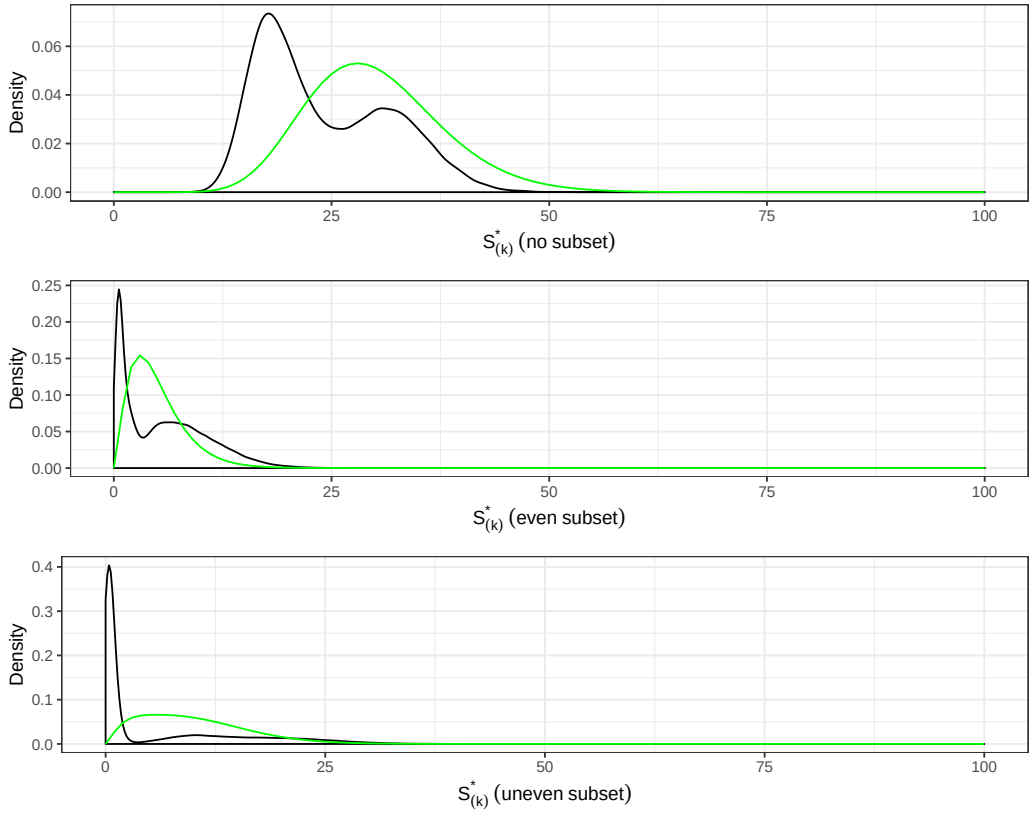


Figure 6.4: PDM density for data set 3

the procedure detailed in Section 2.7.2, and fit the models to the mean hoki catch weight within the 38 grids, for years 2000 – 2008.

Three models were fitted to the gridded hoki catch weight data of the form given in Equations 2.26 – 2.28. We let $\boldsymbol{y}_t = (y(\boldsymbol{g}_1, t), ..., y(\boldsymbol{g}_{38}, t))'$ where $y(\boldsymbol{g}_i, t)$ denoted the log-transformed weighted mean catch weight of hoki in grid $\boldsymbol{g}_i$ for year t, and $n = 38$. The marginal distribution of $\boldsymbol{y}_t$ given the parameters is,

$$\boldsymbol{y}_t|\boldsymbol{\theta} \sim \text{MVN}\Big(\boldsymbol{\mu}_t, \frac{\sigma^2}{1-\rho^2}\boldsymbol{R} + \tau^2\boldsymbol{I}_n\Big),$$

where $\boldsymbol{\mu_t} = \mathbf{1}_{38}\beta$ for each model. The three models are distinguished by the correlation structure assumed for $\boldsymbol{R}$. For model M1, we let the spatial correlation matrix, $\boldsymbol{R}$, be defined by the exponential correlation function, given by Equation 6.6. For model M2, we let the spatial let the spatial correlation matrix, $\boldsymbol{R}$, be defined by the Gaussian correlation function, given by Equation 6.7. For model M3, the spatial correlation matrix, $\boldsymbol{R}$ is defined by the more general Matèrn correlation function,

$$(\boldsymbol{R})_{ij} = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{8\nu}}{\psi}||\boldsymbol{s}_i - \boldsymbol{s}_j||\right)^{\nu} K_\nu\left(\frac{\sqrt{8\nu}}{\phi}||\boldsymbol{s}_i - \boldsymbol{s}_j||\right).$$

For each model, the parameters were assumed *a priori* independent, and were assigned the non-informative prior distributions,

$$\beta \sim \text{N}(0, 100),\ \tau^2 \sim \text{IG}(2, 1),\ \sigma^2 \sim \text{IG}(2, 1),\ \phi \sim \text{U}(0.001, 2),\ \rho \sim \text{U}(-1, 1).$$

Further, for model M3, the smoothness parameter ν was assumed *a priori* independent of the other parameters and was assigned the non-informative prior distribution, $\nu \sim \text{U}(0.01, 10)$.

MCMC was used to fit the three models to the gridded hoki data and this was done through **R** using the package `NIMBLE` (NIMBLE Development Team (2017)). Two chains, each 1000000 iterations, were generated of the parameter vector $\boldsymbol{\theta} = (\beta, \rho, \phi, \sigma^2, \tau^2)'$ for models M1 and M2, and $\boldsymbol{\theta} = (\beta, \rho, \phi, \sigma^2, \tau^2, \nu)'$ for model M3. The first 900000 iterations from each chain were discarded as warm-up, and the remaining draws were combined, resulting in a posterior sample of size $L = 200000$. Table 6.3 gives

Table 6.3: Summary statistics for the models fitted to the hoki data.

	Model 1		Model 2		Model 3	
	Median	(95% CI)	Median	(95% CI)	Median	(95% CI)
β	0.01688	(-0.1793, 0.2083)	0.01815	(-0.1796, 0.2134)	0.01666	(-0.1867, 0.2085)
ν					0.5156	(0.2943, 0.8179)
ϕ	849.1	(584.3, 1000)	261.8	(209.4, 312.9)	828.9	(494.3, 1000)
ρ	0.9675	(0.9381, 0.9879)	0.9899	(0.9776, 0.9978)	0.9687	(0.9347, 0.9913)
τ^2	0.7939	(0.6585, 0.9365)	0.8755	(0.7371, 1.031)	0.7943	(0.6594, 0.9415)
σ^2	0.3106	(0.1470, 0.5463)	0.2408	(0.1287, 0.3944)	0.3082	(0.1435, 0.5365)

Table 6.4: The 10th and 90th percentiles of ordered pivotal discrepancy measures for each model applied to the hoki catch weight data. These are compared to the nominal 10th and 90th percentiles, 0.133 and 28.6 respectively.

nominal percentiles	Model 1	Model 2	Model 3
(0.133, 28.6)	(5.30, 15.8)	(2.90, 13.3)	(5.10, 16.0)

the summary statistics for the posterior parameter distributions computed for each model. There appears to be agreement between the three models on the values of most of the parameters. For model M3, the smoothness parameter ν needed for the Matèrn correlation structure has a posterior median of 0.5156, which means it is close to being estimated as an exponential correlation structure.

For each fitted model, pivotal quantities for every posterior sample were calculated inline with equation 6.2. The grid locations were partitioned into five subsets using the K-means clustering algorithm. The optimal number of subsets to use in the K-means algorithm was found using the elbow method Kodinariya & Makwana (2013). Pivotal quantities for each fitted model $S(\boldsymbol{y}_{tj}, \tilde{\boldsymbol{\theta}}^{(l)})$ for $t = 1, ..., 5$, $j = 1, ..., 5$, and $l = 1, ..., 200000$ were calculated, combined and ordered.

The nominal 10th and 90th percentiles were calculated according to Equations 6.4 and 6.5. They were found to be 0.133 and 28.6 respectively. The

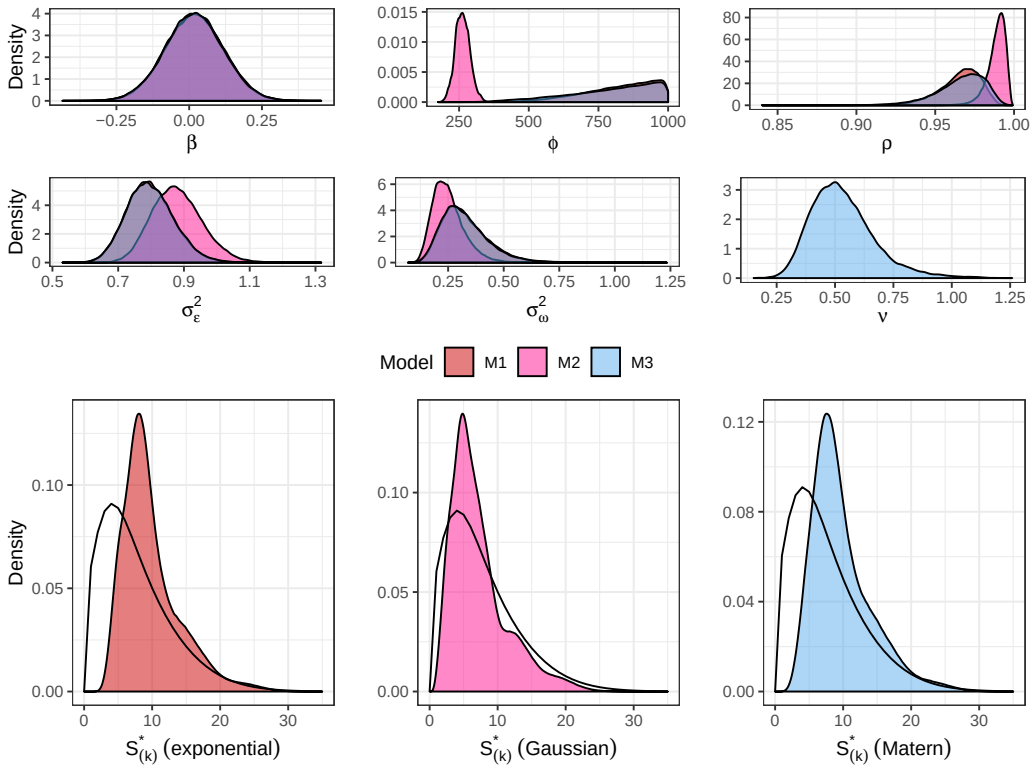


Figure 6.5: Posterior densities for each parameter for each of the three models (top), density of observed ordered PDMs and their nominal distributions for each of the three models (bottom).

10th and 90th percentiles of the ordered pivotal discrepancy measures for models M1, M2 and M3 are given in Table 6.4. It was found that for each model, the 10th percentile was higher than that of the nominal distribution and the 90th percentile was lower than that of the nominal distribution. As a result, it can be said that each model provided a good fit the gridded hoki data. This is reflected in Figure 6. Figure 6 shows the posterior densities for each parameter for each of the three models, as well as the density of the ordered pivotal discrepancy measures. The latter are overlaid with their nominal distributions density. The posterior densities for each parameter are similar, with model M2 having the most different posterior parameter densities. It can be seen that the posterior densities for β, ρ, τ^2, and σ^2 are similar in shape, and overlap between models. For the parameter ϕ, there is a difference between models M1 and M3, and M2. We conclude that Models M1 and M3 are very similar models, and that all three models are providing a similar fit to the data, with the only differences being due to the correlation structure parameters. Looking at the densities of the observed pivotal discrepancy measures for each model, the conclusion that each model provides a good fit is motivated. There is sufficient overlap of the pivotal discrepancy measure densities and their nominal densities such that the test cannot detect a difference.

6.4 Conclusion

In this chapter, we showed that partitioning was necessary when calculating pivotal discrepancy measures for goodness-of-fit. Further, the number of observations within a partition need not be constant. We found that the goodness-of-fit test based on pivotal discrepancy measures was unable to correctly identify an incorrect model misspecification when there was no partitioning. When the number of observations within a partition was not constant, the distribution of the ordered pivotal discrepancy measures were wider, making it more difficult to reject the null hypothesis that the

model provided a good fit, and therefore results in a decrease in power.

Chapter 7

Discussion & concluding remarks

The aim of this research was to develop a range of new methodologies that are capable of accounting for non-stationarity in spatial and spatio-temporal point referenced data. Further, we sought to compare these methodologies to existing ones, in their abilities to make accurate predictions, as well as account for spatial autocorrelation. In this thesis, we contributed three distinct methodologies. In Chapter 3, we proposed partitioned geostatistical models for both spatial and spatio-temporal point referenced data, using the K-means clustering algorithm. In Chapter 4, we proposed covariance regression network models for spatial and spatio-temporal point referenced data. In Chapter 5, we proposed a geographic random forest approach that involved the construction of a neighbourhood structure based on the K-means clustering algorithm. We developed the geographic random forest in both a spatial and spatio-temporal context. Finally, in Chapter 6, we proposed an extension of the pivotal discrepancy measure methodology for Bayesian goodness-of-fit to the spatio-temporal geostatistical model case.

In this chapter, we provide a discussion of the main results from each chapter of this thesis. We then provide a section that compares and contrasts the results from the three methodologies applied to the New Zealand particulate matter data set described in Chapter 2. In addition, we make com-

ments on future areas of research.

7.1 Partitioned geostatistical models

The K-means partitioned geostatistical models that we proposed in Chapter 3 provide a relatively easy and quick method for accounting for non-stationarity while still allowing the use of simple stationary covariance functions. This was highlighted in the simulation studies, particularly in the spatial case. In the spatial simulation, the K-means partitioned geostatistical models (Models 2 – 9) generally provided better predictive accuracy (in terms of RMSE and MAE) when fitted to either stationary or non-stationary point referenced data.

When the simulated data had a stationary spatial structure, we saw lower RMSE and MAE when we assumed global parameters within each sub-region compared to when we assumed local parameters. Furthermore, as the number of partitions increased, RMSE and MAE tended to decrease. This suggested that when we increased the number of partitions and imposed the same covariance structure on each sub-region, then predictive accuracy is improved. However, this was not the case when the parameters were assumed different within each sub-region. Compared to the Matèrn geostatistical model (Model 1), the partitioned models that assumed global coefficients (Models 2 – 5) provided better predictive accuracy in terms of RMSE and MAE, when data has a stationary spatial structure.

When the simulated data had a non-stationary spatial structure, we again saw better predictive accuracy when we assumed global parameters within each sub-region. However, when we fitted the model used to generate the data (Model 7), which assumed local parameters, we saw just as good predictive accuracy. Once more, the partitioned models that assumed global coefficients (Models 2 – 5), and Model 7, which correctly specified the spatial structure of the simulated data, provided better predictive accuracy in

terms of RMSE and MAE, when data has a non-stationary spatial structure.

For both sets of simulated spatial data, the models that assumed local parameters within each sub-region accounted for less spatial autocorrelation compared to the models that assumed global paramters. This was reflected by the values of Moran's I calculated on the residuals, and suggests that when using partitioned geostatistical models there might be a trade-off between predictive accuracy and accounting for spatial autocorrelation.

The results of the spatial simulation were reinforced when we fitted partitioned geostatistical models to the New Zealand particulate matter data observed in 2013. We found that when we partitioned the particulate matter monitoring stations into two or three sub-regions and and fitted models that assumed local parameters within each sub-region (Models 4 and 5), we found better predictive accuracy. Unlike the simulation results, we also found that these models were able to account for the most spatial autocorrelation out of the five that we fitted. These results show that there is great potential for fitting K-means partitioned geostatistical models to spatial point referenced data.

Results for the stationary spatio-temporal simulation somewhat deviated from that of the stationary spatial simulation. We found that Model 5 provided the best predictive accuracy, which was based on five partitions and assumed global coefficients within each sub-region. This model performed better, in terms of predictive accuracy, than the traditional Matèrn geostatistical model, which was also the model that generated the data. However, this might be due to the lack of convergence for some of the parameter posterior distributions. Unfortunately, due to computational restrictions, we were not able to run the model for a larger number of iterations. A similar set of results were observed for the non-stationary spatio-temporal simulation.

While these results show potential for the partitioned geostatistical model,

we believe there is room for future considerations. Firstly, a repeat of the simulation study using different covariance structures to simulate the data could be performed. This would allow for an investigation into the effect that different covariance structures have on partitioned geostatistical model fitting, in terms of predictive accuracy and spatial autocorrelation. Furthermore, we would run the models for longer, using a larger number of iterations in the simulations and case studies. This would lessen the influence of posterior distributions that had not fully converged, which was the case in the spatio-temporal simulation and case studies.

In addition, we would consider different partitioning algorithms, including those that might account for non-stationarity in different ways. One issue with the K-means approach to partitioning involves the choice of selecting K, the number of partitions. We propose that Bayesian model averaging could be used to eliminate the subjectivity of selecting K.

Another consideration is that of observations located at the boundaries of the partitions. The models that we fitted assumed that each sub-region was independent of each other. This meant that two observations that were separated by a short distance, but assigned to different sub-regions by the K-means algorithm, were assumed to be uncorrelated. In reality, this is unlikely to be the case. In future, we would like to explore the possibility of allowing boundary effects.

We might also develop a model that allows for modelling spatio-temporal data that is misaligned across the time. The models that we described in Chapter 3 assumed that the locations at which observations were made do not change over time. A better modelling framework would allow for spatio-temporal data that has been observed at locations that change over time. The sub-Antarctic hoki dataset described in Chapter 2 is an example of this type of spatio-temporal data. We attempted a case study on the gridded hoki data, but found that partitioning the data meant that there were too few observations within each sub-region. As a result, we did not fit partitioned geostatistical models to that data.

7.2 Covariance regression network models

Covariance regression network (CRN) models were proposed for spatial and spatio-temporal point referenced data in Chapter 4. They were shown to provide more flexibility in modelling the covariance function of spatial and spatio-temporal processes. The best results in terms of predictive accuracy measures, RMSE and MAE, were obtained when we performed Bayesian model averaging over the CRN models that were based on different estimates of the network structure.

Once again, however, we found a trade-off between predictive accuracy and accounting for spatial autocorrelation. In general, we observed an increasing trend in Moran's I as the network structure grew to include more locations.

A disadvantage to modelling the covariance function of spatial and spatio-temporal processes using covariance regression network models is that the network structure of the point referenced is almost always unknown. Within the field of network analysis, from which CRN models were conceived, the network structure of locations is known. For spatial point referenced data, the network structure is latent and needs to be estimated. We estimated the network structure using the adjacency matrix constructed via a distance function. Future considerations for CRN models applied to spatial data might include other ways of estimating the network structure.

Another issue that was encountered when fitting CRN models was that of the positive definiteness condition for the covariance matrix. Lan et al. (2018) imposed constraints on the range of values for the regression coefficients so that when they were estimated (in a frequentist sense), the resulting covariance matrix would be positive definite. We used the Bayesian framework to estimate the regression coefficients and as such, did not impose constraints. This lead to issues with some draws from the posterior distributions of the regression coefficients leading to estimates of the covariance matrix that were not positive definite. In future, we would con-

sider imposing constraints in the form of prior distributions on the regression coefficients.
The CRN models that we fitted to spatio-temporal point referenced data assumed that the neighbourhood structure did not change over time. To add another layer of flexibility to this methodology, we would consider allowing the network structure to change over time, to allow for capturing spatio-temporal autocorrelation.

7.3 Geographic random forest

In Chapter 5, we proposed an extension of the geographic random forest methodology to incorporate neighbourhood structures that were constructed via K-means clustering. Further, we extended the methodology to the spatio-temporal realm. The benefit to using geographic random forest over the previous two methodologies that we proposed, is the lack of needing to provide a covariance structure. We did not find improvements in predictive accuracy when K-means clustering was used to define the neighbourhood structure.
In future, we would look at developing better ways of defining a neighbourhood structure for spatial and spatio-temporal data. Furthermore, we would look at different ways of defining the structure over time.

7.4 Comparison of methodologies

In Chapter's 3, 4, and 5, we fitted our proposed methodologies and a traditional Matèrn covariance model to the New Zealand particulate matter dataset. This allows us to compare the best performing models (in terms of predictive accuracy) within each proposed methodology and to a traditional method. Table 7.1 displays the median posterior RMSE, MAE, and Moran's I (calculated on the residuals) for the traditional Matèrn covariance model, fitted in Chapter 3, and for the best performing models in

Chapter's 3, 4, and 5. The best performing model in Chapter 3 was the K-means partitioned covariance model with two partitions and assuming local coefficients for each sub-region. The best performing model in Chapter 4 was the Bayesian model averaged covariance regression network model. The best performing model in Chapter 5 was the geographic random forest using a clustering approach.

In table 7.1, we see that the model fitted in Chapter 3, a K-means partitioned covariance model using two partitions and assuming local coefficients for each sub-region performed the best of all three proposed methodologies in terms of predictive accuracy. Further, it performed better than the traditional Matèrn covariance model in terms of predictive accuracy. However, this model did not account for spatial autocorrelation the best. The model that accounted for spatial autocorrelation the best was that of the geographic random forest, because it had the smallest median posterior Moran's I, calculated on the residuals. Each proposed methodology was able to account for more spatial autocorrelation than the traditional Matèrn model.

Table 7.1: Median posterior RMSE, MAE, and Moran's I (calculated on the residuals) for the traditional Matèrn model, and the best performing (in terms of predictive accuracy) models of Chapter 3, 4, and 5, fitted to the New Zealand particulate matter dataset. The traditional Matèrn model was fitted in Chapter 3.

	RMSE	MAE	Moran's I
Traditional	0.902	0.754	0.342
Partitioned	0.655	0.528	0.160
Covariance regression	2.737	2.334	0.116
Geographic random forest	2.600	2.050	0.096

7.5 Pivotal discrepancy measures

From the simulation study, it is clear that partitioning is necessary, and furthermore, that the number of observations within a partition need not be constant. We found that the goodness-of-fit test based on pivotal discrepancy measures was unable to correctly identify an incorrect model misspecification when there was no partitioning. Further, when the number of observations within a partition was not constant, the distribution of the ordered pivotal discrepancy measures were wider, making it more difficult to reject the null hypothesis that the model provided a good fit, and therefore results in an increase in power.

The choice of how to partition the data should be considered carefully. In the simulation study in this paper, the subsets were chosen sensibly, in that we partitioned the observations according to the subsets that were used to generate the data. In reality, this will not necessarily be known, and an objective method should be developed. In the case study, we showed that using the K-means clustering algorithm is a suitable approach to partitioning.

A final consideration is that of how to select the best model from competing models fit to the same data. The goodness-of-fit test based on pivotal discrepancy measures currently offers no way to select the best model, instead opting for a decision based test only. Jun et al. (2014) and Johnson (2007) talk briefly on calculating bounds on Bayesian p-values that may offer an appropriate route to model selection.

In conclusion, we have developed a general goodness-of-fit test for Bayesian spatio-temporal models using partitioning and pivotal discrepancy measures. It has seen success in simulation as well as application to New Zealand hoki data.

This thesis investigated and developed a range of non-parametric techniques to flexibly model spatial and spatio-temporal point referenced data, while accounting for non-stationarity.

Appendix A

Convergence diagnostics for Chapter 3

For conciseness, diagnostic plots for convergence are only supplied for one model applied to stationary and non-stationary data for one repetition of the simulation. The remaining plots can be found at `https://github.com/morrislind/PhD2020`.

Table A.1: Potential scale reduction factor calculated for each parameter for Models 1 – 5 fitted to the first repetition of simulated spatial stationary data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	1.000	1.041	1.010	1.004	1.011
β_1	1.007	0.995	0.996	0.998	0.997
ψ	0.998	1.131	1.004	1.097	0.996
σ^2	0.996	1.040	1.004	0.997	1.000
τ^2	1.004	1.005	0.995	1.030	1.022

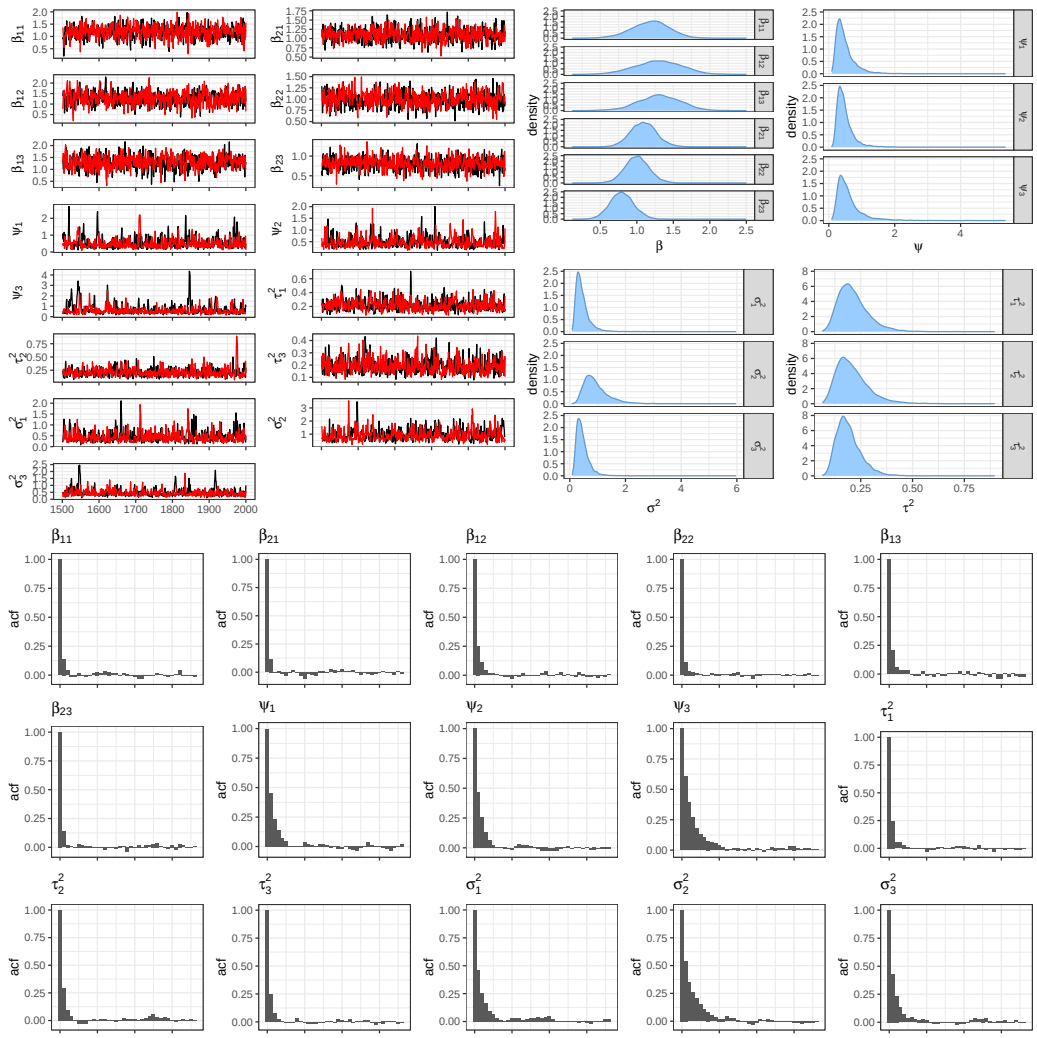


Figure A.1: Convergence diagnostic plots for Model 7 fitted to the first set of stationary spatial simulated data. Convergence to stationary posterior distributions is satisfied.

Table A.2: Potential scale reduction factor calculated for each parameter for Models 6 – 9 fitted to the first repetition of simulated spatial stationary data.

Parameter	Model 6	Model 7	Model 8	Model 9
β_{01}	0.995	1.018	0.995	0.997
β_{11}	1	0.998	1.004	0.998
β_{02}	1.033	1.002	1.022	1.012
β_{12}	0.998	1.078	0.995	1.01
β_{03}		1.008	0.996	1.008
β_{13}		0.996	0.996	0.995
β_{04}			0.995	1.011
β_{14}			0.995	1.001
β_{05}				1.006
β_{15}				1.046
ψ_1	0.998	1.014	1.007	1.141
ψ_2	0.995	1.022	1.012	1.022
ψ_3		1.215	0.995	1.018
ψ_4			1.026	1.031
ψ_5				1.056
σ_1^2	0.995	0.995	1.004	0.995
σ_2^2	1.023	1.003	1.003	0.996
σ_3^2		1.064	1.043	1.02
σ_4^2			0.999	1.008
σ_5^2				1.049
τ_1^2	0.996	0.997	0.996	1.004
τ_2^2	1.002	0.995	1.039	1.007
τ_3^2		1.008	1.04	1.052
τ_4^2			1.022	0.998
τ_5^2				0.996

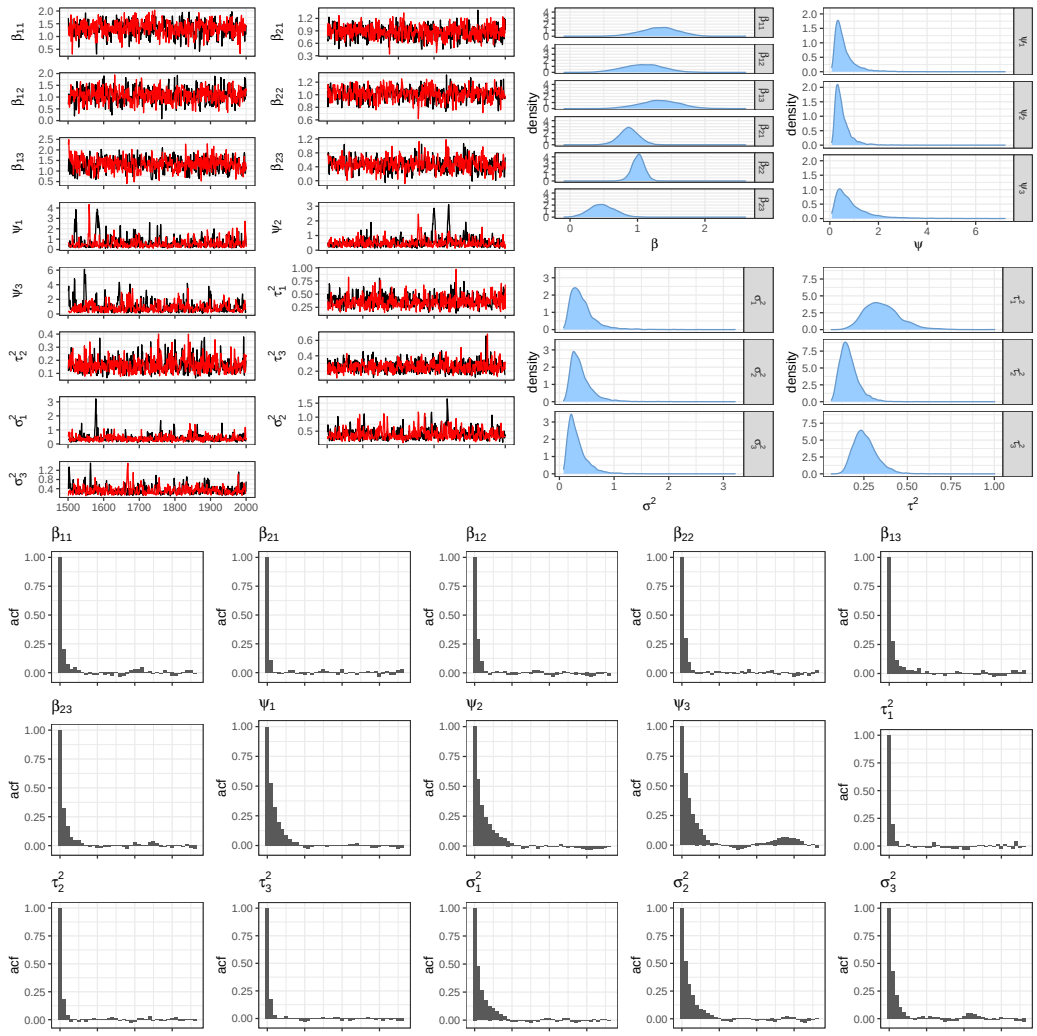


Figure A.2: Convergence diagnostic plots for Model 7 fitted to the first set of non-stationary spatial simulated data. Convergence to stationary posterior distributions is satisfied.

Table A.3: Potential scale reduction factor calculated for each parameter for Models 1 – 5 fitted to the first repetition of simulated spatial non-stationary data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	1.014	1.030	0.995	1.013	0.998
β_1	1.000	1.048	1.000	1.005	1.008
ψ	1.020	0.995	1.116	1.017	1.015
σ^2	0.996	1.016	1.037	1.009	0.999
τ^2	1.003	1.038	0.995	1.087	0.995

Table A.4: Potential scale reduction factor calculated for each parameter for Models 6 – 9 fitted to the first repetition of simulated spatialnon- stationary data.

Parameter	Model 6	Model 7	Model 8	Model 9
β_{01}	0.998	1.024	1.009	0.996
β_{11}	1.015	1.031	0.996	0.998
β_{02}	1.014	0.995	0.995	1.019
β_{12}	1.000	0.997	0.995	1.021
β_{03}		1.006	0.995	1.005
β_{13}		0.995	1.004	0.997
β_{04}			0.996	1.017
β_{14}			0.998	0.996
β_{05}				1.091
β_{15}				1.080
ψ_1	0.999	1.032	1.003	0.997
ψ_2	1.001	1.027	0.996	1.003
ψ_3		1.045	0.998	0.997
ψ_4			1.111	1.067
ψ_5				1.003
σ_1^2	0.996	0.995	0.998	1.004
σ_2^2	1.002	0.995	1.003	0.997
σ_3^2		0.995	1.002	0.995
σ_4^2			0.998	1.001
σ_5^2				0.995
τ_1^2	1.074	1.038	0.996	1.008
τ_2^2	1.008	0.998	0.996	0.997
τ_3^2		0.995	0.996	1.105
τ_4^2			1.023	1.061
τ_5^2				1.008

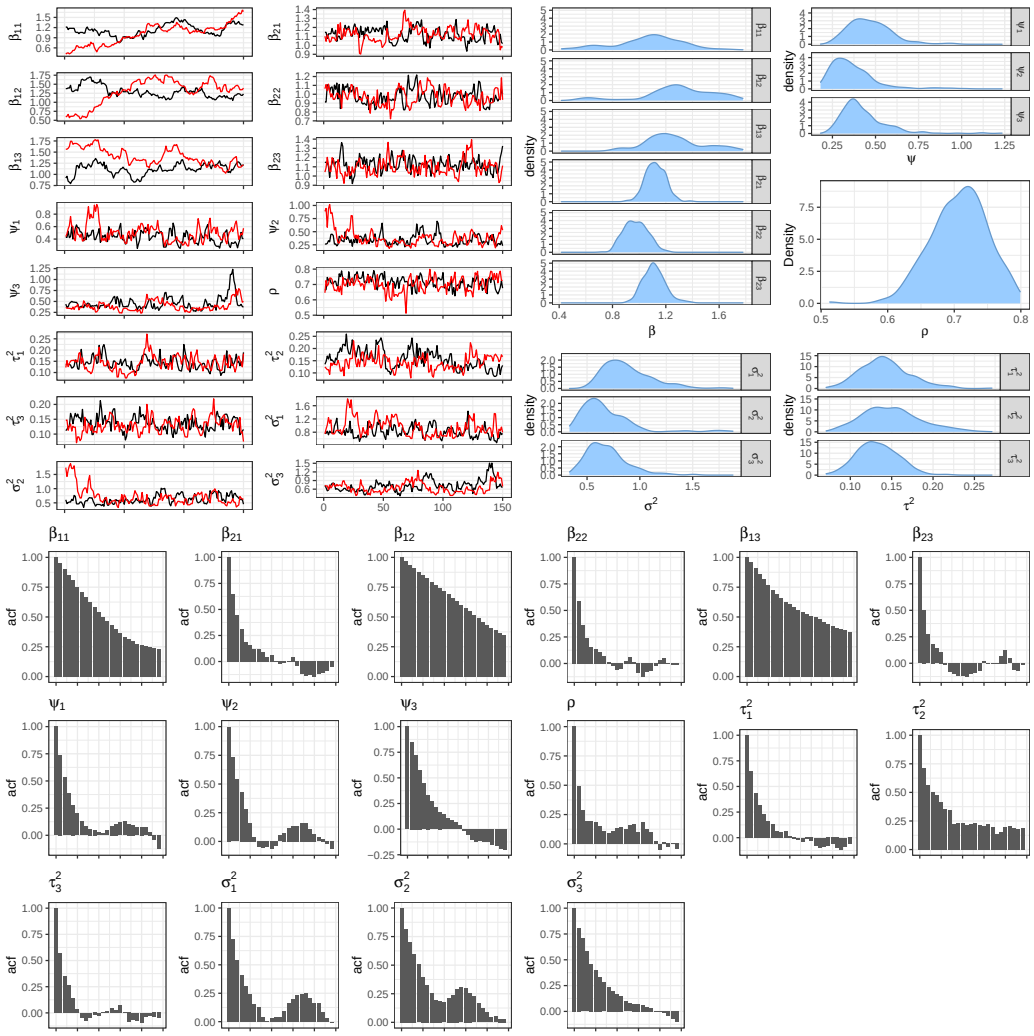


Figure A.3: Convergence diagnostic plots for Model 7 fitted to the first set of stationary spatio-temporal simulated data. There is indication that not all parameters converged to stationary posterior distributions.

Table A.5: Potential scale reduction factor calculated for each parameter for Models 1 – 5 fitted to the first repetition of simulated spatio-temporal stationary data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	4.531	4.464	1.513	1.061	37.636
β_1	1.082	0.997	0.997	0.998	1.105
ψ	0.999	1.013	0.995	1.350	1.660
ρ	1.216	1.001	1.012	1.004	1.431
σ^2	1.077	1.030	0.994	1.146	1.206
τ^2	1.054	1.074	0.993	1.167	1.049

Table A.6: Potential scale reduction factor calculated for each parameter for Models 6 – 9 fitted to the first repetition of simulated spatio-temporal stationary data.

Parameter	Model 6	Model 7	Model 8	Model 9
β_{01}	2.961	1.178	0.995	1.106
β_{11}	1.231	0.993	1.002	1.002
β_{02}	12.612	0.993	2.675	3.557
β_{12}	1.063	0.997	1.042	0.999
β_{03}		3.444	2.496	1.191
β_{13}		1.087	1.408	1.004
β_{04}			2.653	1.237
β_{14}			1.003	0.995
β_{05}				0.994
β_{15}				1.016
ψ_1	2.778	1.214	1.578	0.994
ψ_2	1.217	1.047	1.046	0.997
ψ_3		1.195	1.015	1.005
ψ_4			1.024	1.413
ψ_5				0.995
ρ	1.047	1.078	1.434	0.994
σ_1^2	1.053	0.993	1.087	1.010
σ_2^2	1.427	1.141	1.187	1.049
σ_3^2		1.053	1.007	0.998
σ_4^2			1.029	0.994
σ_5^2				0.993
τ_1^2	2.324	1.267	1.059	0.994
τ_2^2	1.086	1.083	1.172	1.051
τ_3^2		1.100	1.005	1.081
τ_4^2			1.100	1.225
τ_5^2				0.998

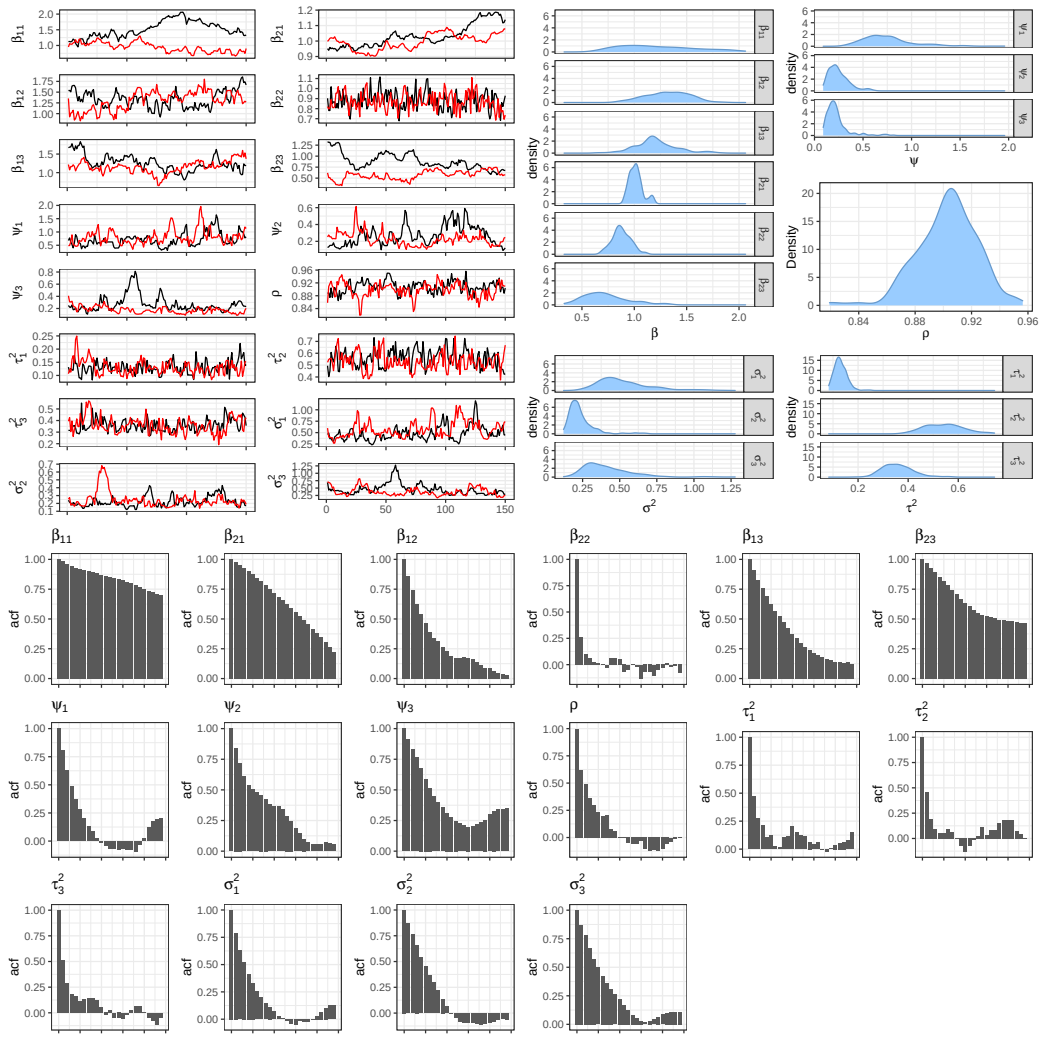


Figure A.4: Convergence diagnostic plots for Model 7 fitted to the first set of non-stationary spatio-temporal simulated data. There is indication that not all parameters converged to stationary posterior distributions.

Table A.7: Potential scale reduction factor calculated for each parameter for Models 1 – 5 fitted to the first repetition of simulated spatio-temporal non-stationary data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	4.531	4.464	1.513	1.061	37.636
β_1	1.082	0.997	0.997	0.998	1.105
ψ	0.999	1.013	0.995	1.350	1.660
ρ	1.216	1.001	1.012	1.004	1.431
σ^2	1.077	1.030	0.994	1.146	1.206
τ^2	1.054	1.074	0.993	1.167	1.049

Table A.8: Potential scale reduction factor calculated for each parameter for Models 6 – 9 fitted to the first repetition of simulated spatio-temporal non-stationary data.

Parameter	Model 6	Model 7	Model 8	Model 9
β_{01}	2.961	1.178	0.995	1.106
β_{11}	1.231	0.993	1.002	1.002
β_{02}	12.612	0.993	2.675	3.557
β_{12}	1.063	0.997	1.042	0.999
β_{03}		3.444	2.496	1.191
β_{13}		1.087	1.408	1.004
β_{04}			2.653	1.237
β_{14}			1.003	0.995
β_{05}				0.994
β_{15}				1.016
ψ_1	2.778	1.214	1.578	0.994
ψ_2	1.217	1.047	1.046	0.997
ψ_3		1.195	1.015	1.005
ψ_4			1.024	1.413
ψ_5				0.995
ρ	1.047	1.078	1.434	0.994
σ_1^2	1.053	0.993	1.087	1.010
σ_2^2	1.427	1.141	1.187	1.049
σ_3^2		1.053	1.007	0.998
σ_4^2			1.029	0.994
σ_5^2				0.993
τ_1^2	2.324	1.267	1.059	0.994
τ_2^2	1.086	1.083	1.172	1.051
τ_3^2		1.100	1.005	1.081
τ_4^2			1.100	1.225
τ_5^2				0.998

Appendix B

Convergence diagnostics for Chapter 4

For conciseness, diagnostic plots for convergence are only supplied for one model fitted in each simulation. The remaining plots can be found at `https://github.com/morrislind/PhD2020`.

Table B.1: Potential scale reduction factor calculated for each parameter of the CRN Models 1 – 10 and Matèrn Model 11 fitted to the first repetition of simulated spatial data.

Model	β	γ_0	γ_1	γ_2
1	0.998	0.997	0.999	1.004
2	1.009	1.081	1.063	1.043
3	1.008	0.997	1.038	1.001
4	1.007	1.092	1.082	1.000
5	1.004	1.014	1.000	1.003
6	1.075	1.094	1.027	0.998
7	1.178	1.053	0.998	0.999
8	0.997	1.004	1.031	0.998
9	1.061	1.047	1.080	0.998
10	0.998	1.018	1.044	0.999
Model	β	ψ	σ^2	τ^2
11	0.998	1.191	1.000	1.218

Table B.2: Potential scale reduction factor calculated for each parameter of the CRN Models 1 – 10 and Matèrn Model 11 fitted to the first repetition of simulated spatio-temporal data.

Model	β_{01}	β_{02}	γ_{01}	γ_{02}	γ_{11}	γ_{12}	γ_{21}	γ_{22}
1	1.000	1.067	1.000	1.090	1.018	1.029	1.000	1.053
2	1.103	1.005	1.002	1.003	1.092	0.997	1.013	1.024
3	0.997	1.018	1.003	1.009	1.012	1.004	1.011	1.001
4	1.000	1.040	1.014	1.023	1.008	1.032	0.999	1.079
5	0.999	1.043	1.035	0.998	1.050	1.091	0.997	1.054
6	1.042	1.013	1.012	1.046	1.001	1.052	0.997	1.021
7	1.015	1.026	1.063	1.064	1.024	1.003	0.999	1.019
8	1.039	0.997	1.066	1.085	1.013	1.127	1.010	1.013
9	1.042	1.182	1.221	1.163	1.052	1.021	1.016	0.998
10	1.147	1.022	1.064	1.015	0.998	0.998	0.999	0.999
Model	β_{01}	β_{02}	ψ	ρ	σ^2	τ^2		
11	1.033	0.999	0.997	1.022	1.110	1.011		

Table B.3: Potential scale reduction factor calculated for each parameter of Models 1 – 10 fitted to the NZ PM10 concentration data.

Model	β_0	β_1	β_2	γ_0	γ_1	γ_2
1	1.000	1.001	1.001	1.023	1.000	1.007
2	1.000	1.010	1.007	1.017	1.101	1.134
3	1.001	1.006	1.003	1.022	1.036	1.031
4	1.000	1.000	1.000	1.001	1.021	1.061
5	1.006	1.007	1.000	1.380	1.580	1.364
6	1.000	1.004	1.000	1.053	1.050	1.021
7	1.001	1.003	1.001	1.097	1.046	1.017
8	1.001	1.010	1.000	1.014	1.120	1.157
9	1.001	1.012	1.001	1.089	1.125	1.128
10	1.027	1.211	1.795	1.397	1.344	1.238

Table B.4: Potential scale reduction factor calculated for the β parameters for Models 1 – 5 fitted to the hoki catch weight data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_{01}	1.000	1.001	1.000	1.000	1.000
β_{11}	1.000	1.000	1.000	1.005	1.001
β_{02}	1.000	1.002	1.000	1.002	1.000
β_{12}	1.000	1.000	1.001	1.001	1.000
β_{03}	1.002	1.000	1.000	1.000	1.001
β_{13}	1.002	1.001	1.001	1.000	1.001
β_{04}	1.000	1.000	1.000	1.004	1.000
β_{14}	1.000	1.002	1.000	1.000	1.003
β_{05}	1.001	1.000	1.001	1.000	1.001
β_{15}	1.039	1.002	1.006	1.000	1.000
β_{06}	1.000	1.000	1.000	1.000	1.001
β_{16}	1.001	1.000	1.003	1.000	1.002
β_{07}	1.000	1.000	1.005	1.000	1.000
β_{17}	1.002	1.000	1.000	1.000	1.001
β_{08}	1.001	1.000	1.003	1.002	1.004
β_{18}	1.001	1.000	1.000	1.001	1.000
β_{09}	1.002	1.000	1.003	1.000	1.001
β_{19}	1.001	1.000	1.006	1.004	1.002

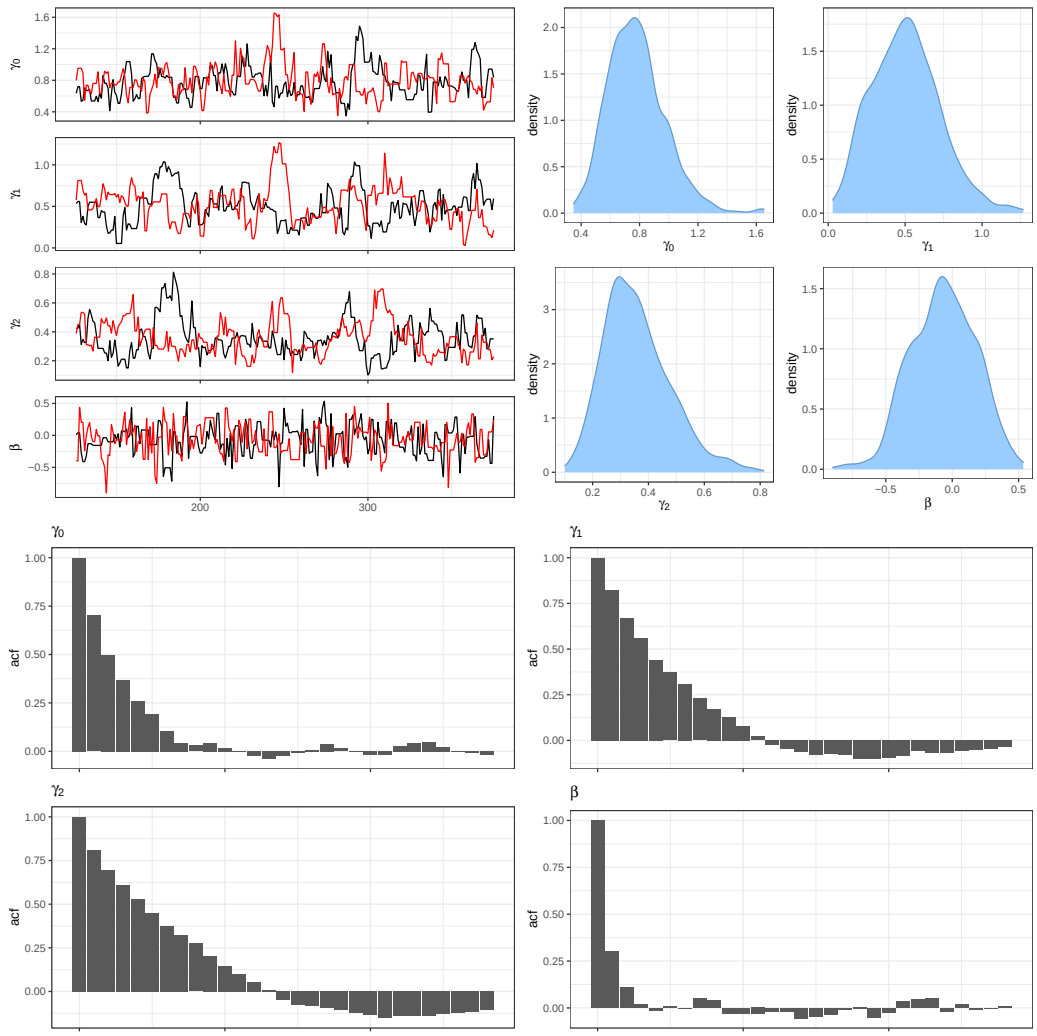


Figure B.1: Convergence diagnostic plots for Model 1 fitted to the first repetition of simulated spatial data. There is indication that all parameters converged to stationary posterior distributions.

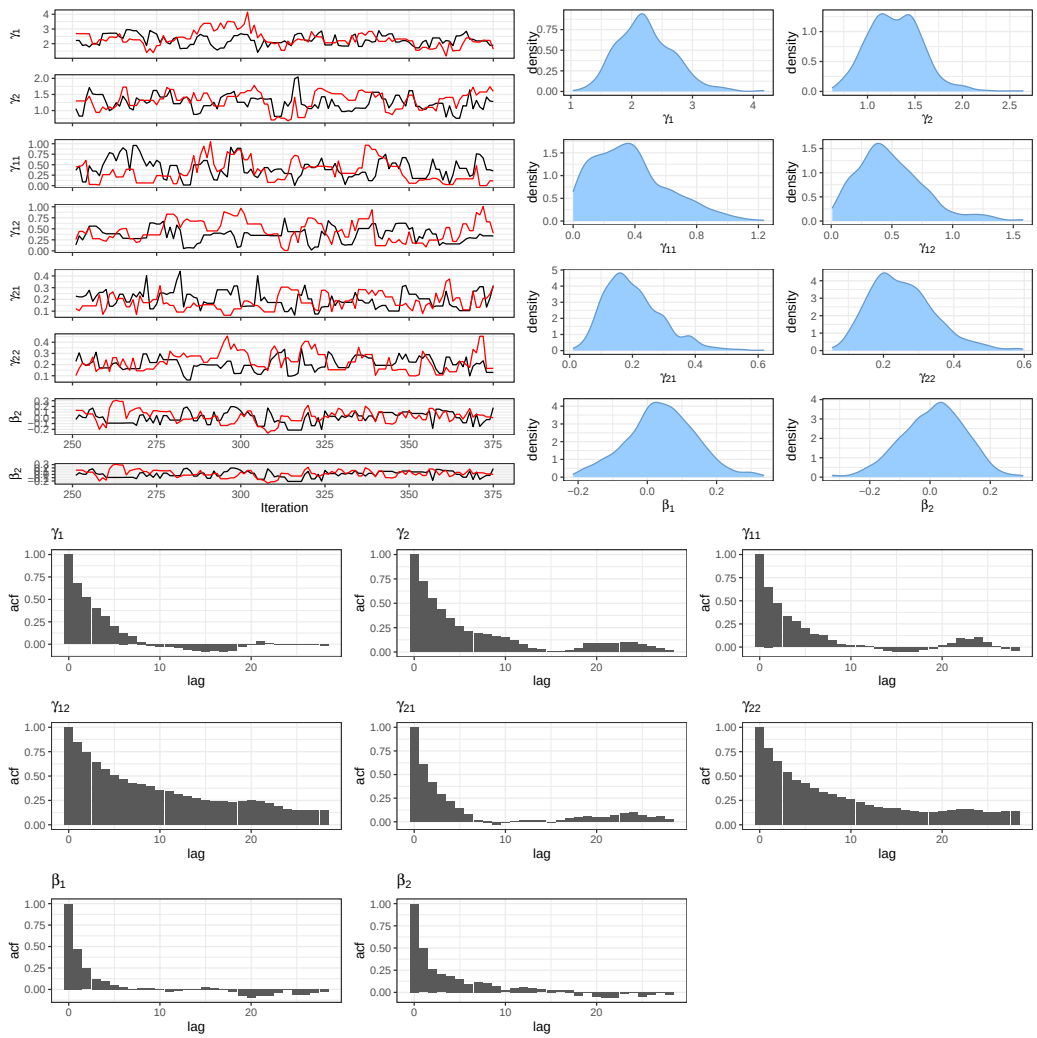


Figure B.2: Convergence diagnostic plots for Model 1 fitted to the first repetition of simulated spatio-temporal data. There is indication that all parameters converged to stationary posterior distributions.

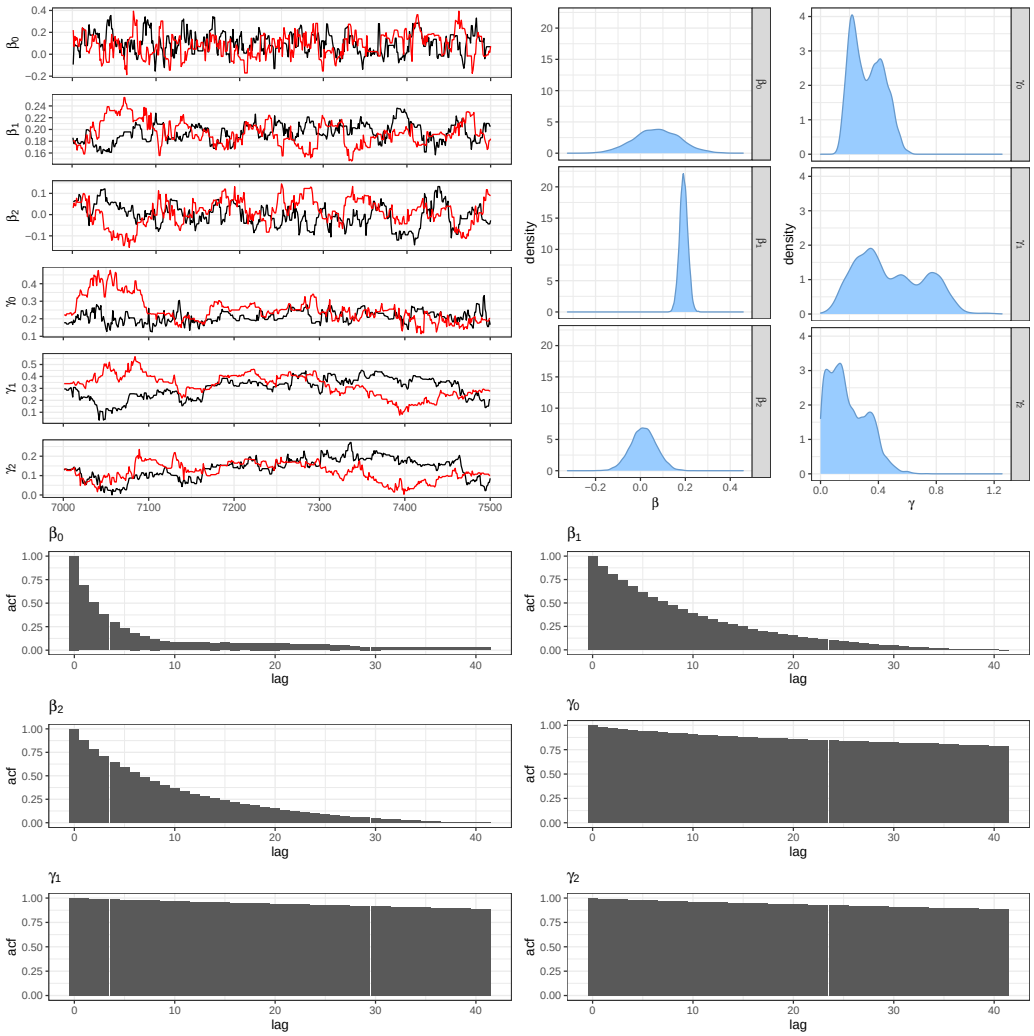


Figure B.3: Convergence diagnostic plots for Model 1 fitted to the NZ PM10 concentration data. There is indication that all parameters converged to stationary posterior distributions.

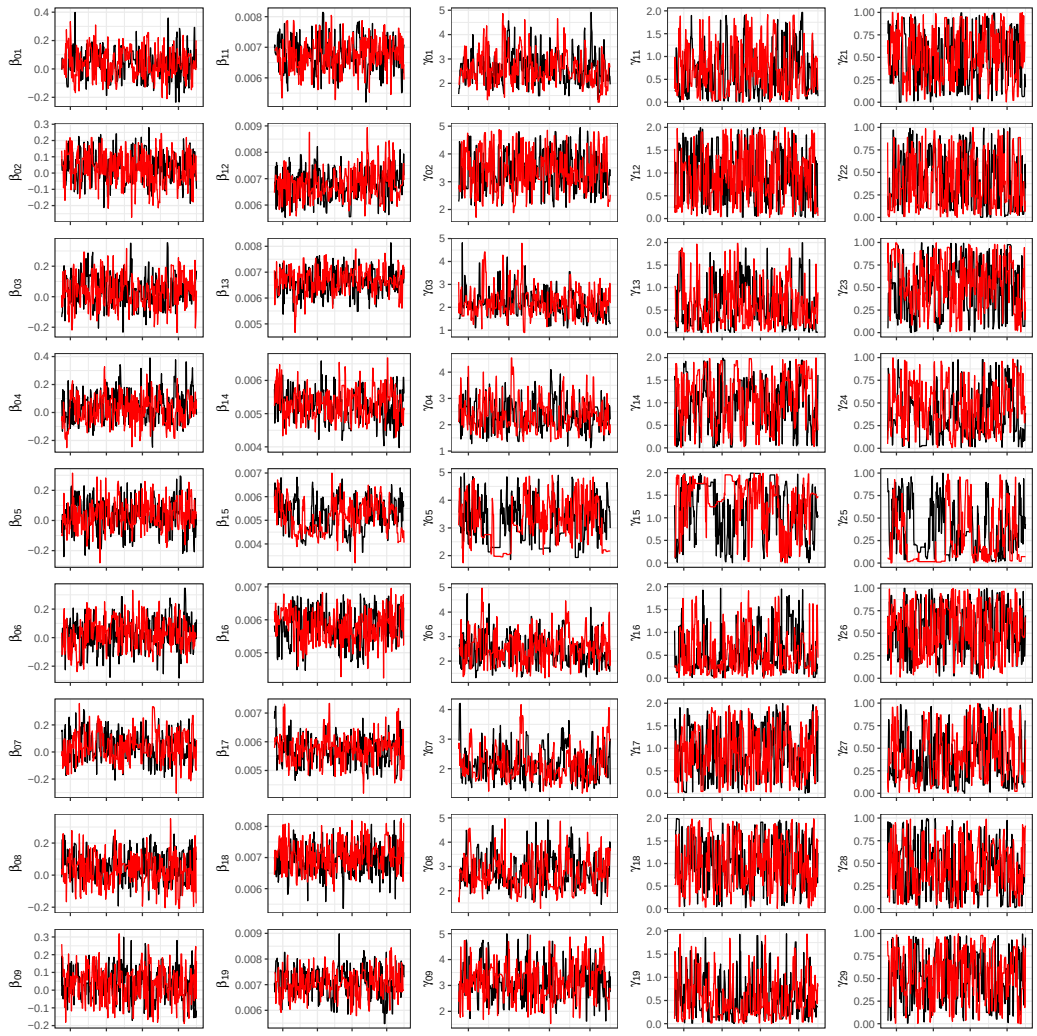


Figure B.4: Trace plots for Model 1 fitted to the hoki catch weight data. There is indication that all parameters converged to stationary posterior distributions.

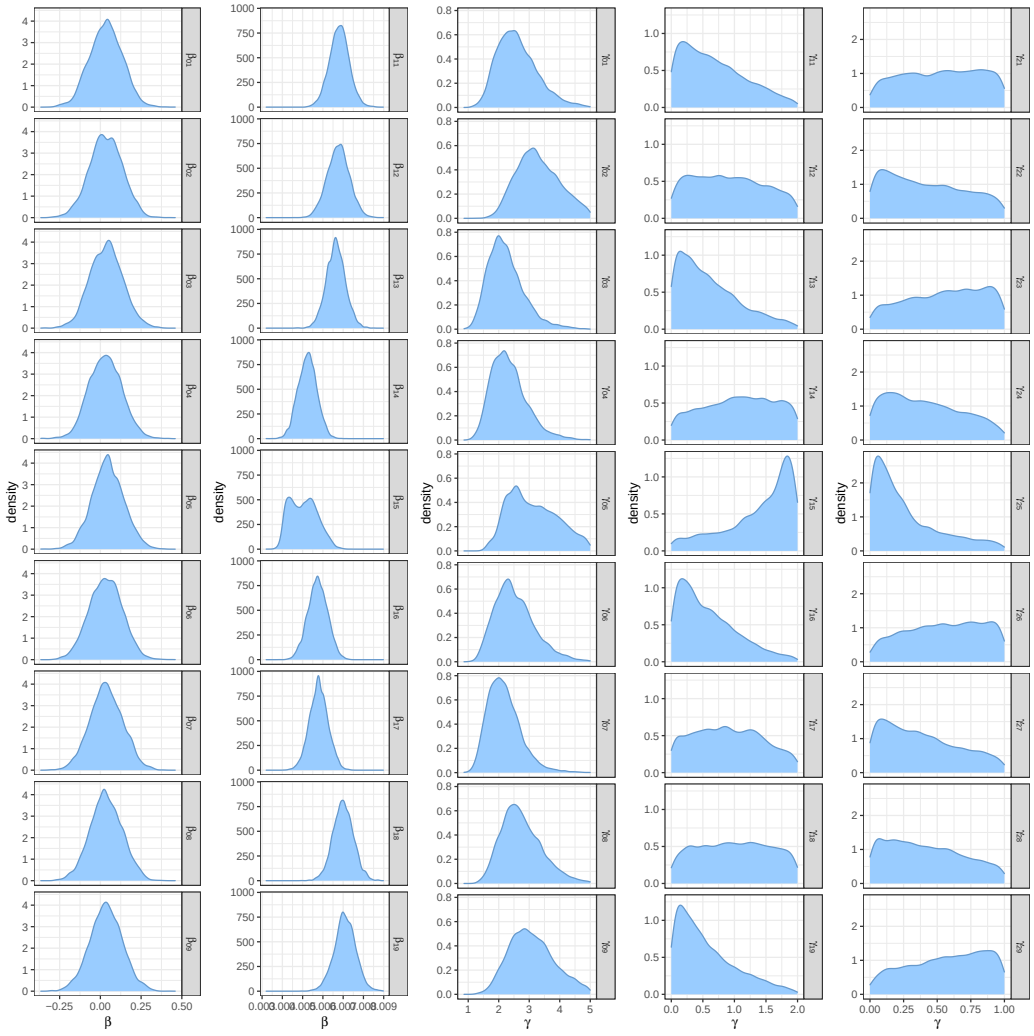


Figure B.5: Density plots for Model 1 fitted to the hoki catch weight data. There is indication that all parameters converged to stationary posterior distributions.

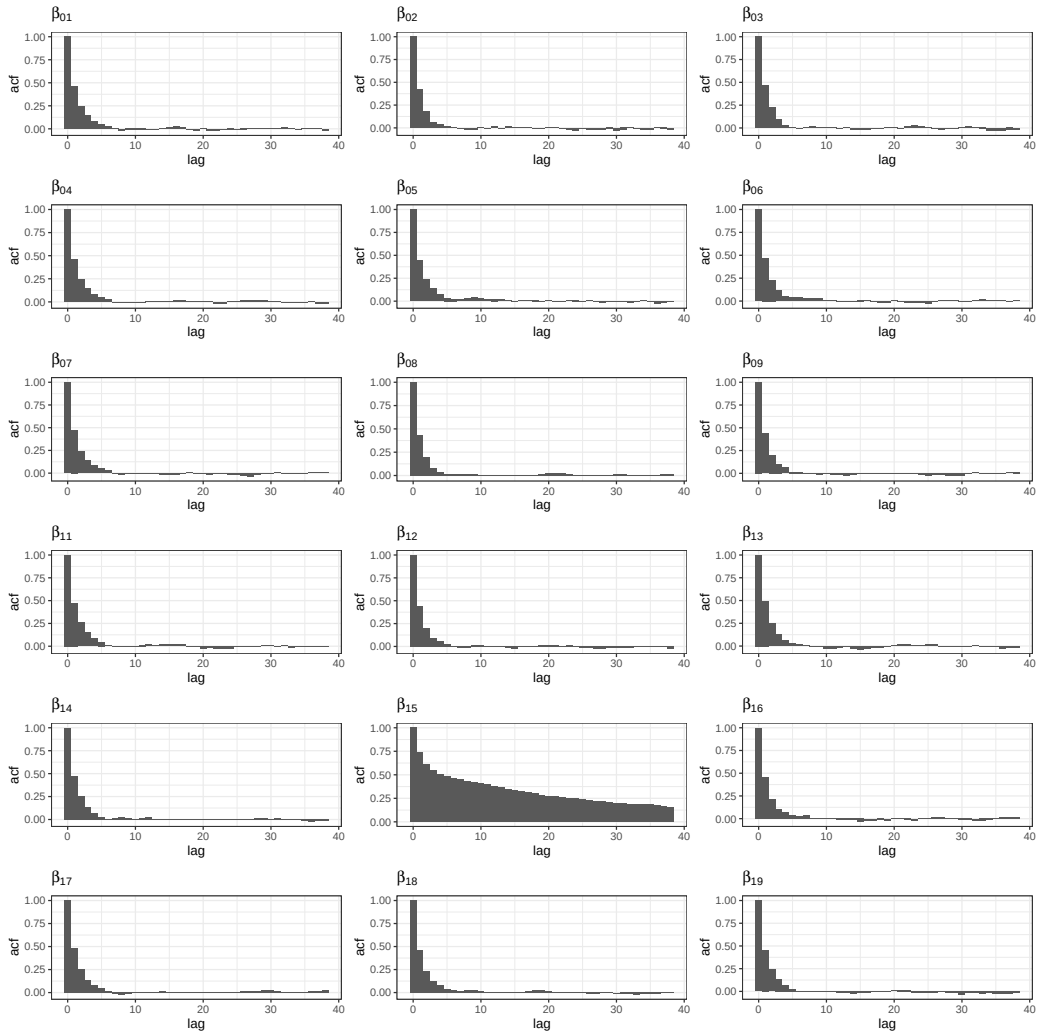


Figure B.6: Autocorrelation function plots for the β parameters of Model 1 fitted to the hoki catch weight data.

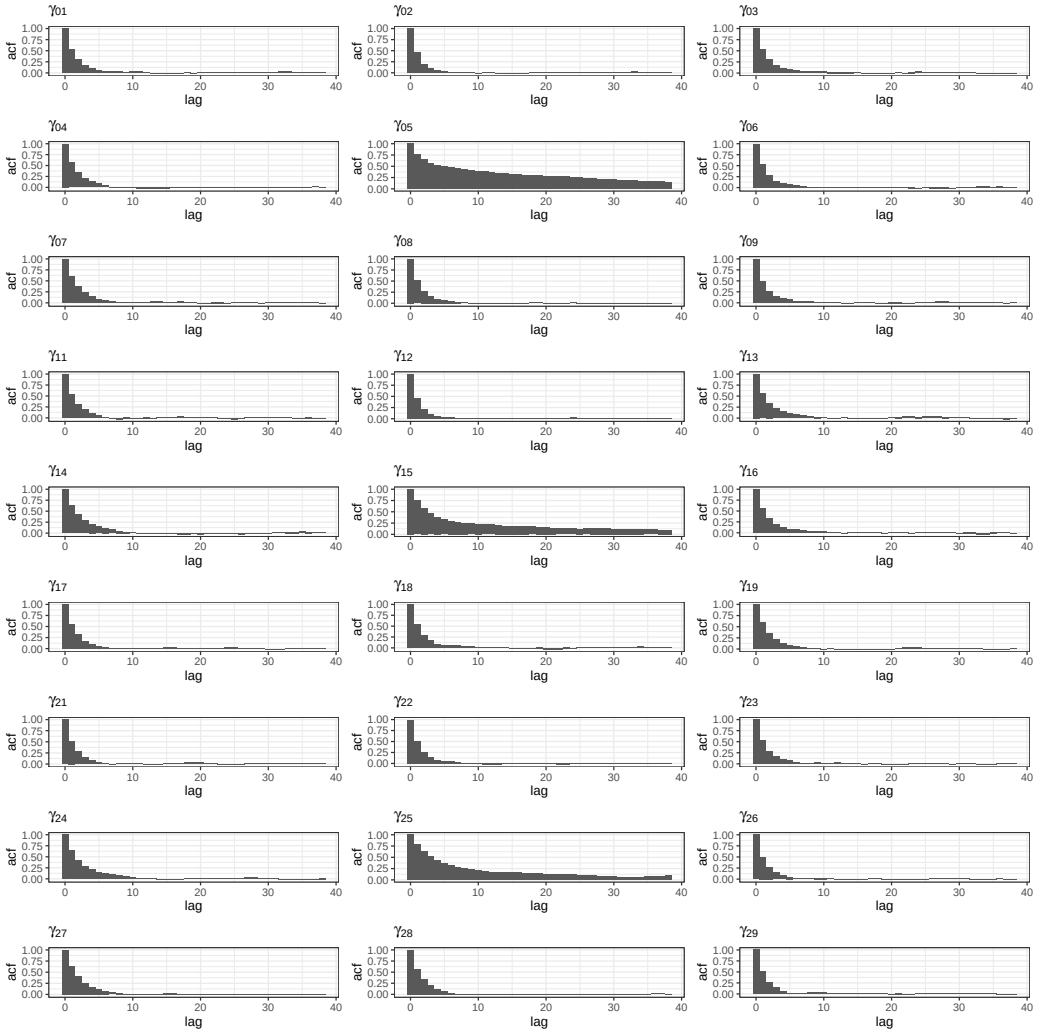


Figure B.7: Autocorrelation function plots for the γ parameters of Model 1 fitted to the hoki catch weight data.

Table B.5: Potential scale reduction factor calculated for the γ parameters for Models 1 – 5 fitted to the hoki catch weight data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
γ_{01}	1.004	1.000	1.000	1.000	1.002
γ_{11}	1.000	1.000	1.000	1.002	1.000
γ_{21}	1.005	1.008	1.005	1.000	1.005
γ_{02}	1.001	1.001	1.000	1.004	1.008
γ_{12}	1.000	1.001	1.002	1.000	1.004
γ_{22}	1.000	1.001	1.000	1.001	1.003
γ_{03}	1.002	1.000	1.003	1.000	1.001
γ_{13}	1.000	1.000	1.000	1.001	1.000
γ_{23}	1.000	1.000	1.000	1.005	1.000
γ_{04}	1.003	1.006	1.020	1.002	1.004
γ_{14}	1.000	1.006	1.000	1.002	1.000
γ_{24}	1.000	1.002	1.002	1.001	1.003
γ_{05}	1.058	1.000	1.000	1.002	1.000
γ_{15}	1.002	1.002	1.001	1.002	1.000
γ_{25}	1.027	1.000	1.004	1.004	1.000
γ_{06}	1.000	1.003	1.003	1.000	1.005
γ_{16}	1.000	1.007	1.001	1.002	1.003
γ_{26}	1.000	1.000	1.002	1.001	1.001
γ_{07}	1.000	1.013	1.002	1.020	1.003
γ_{17}	1.000	1.001	1.007	1.037	1.000
γ_{27}	1.000	1.006	1.004	1.004	1.007
γ_{08}	1.003	1.000	1.006	1.003	1.000
γ_{18}	1.006	1.000	1.000	1.000	1.000
γ_{28}	1.012	1.000	1.009	1.005	1.001
γ_{09}	1.000	1.001	1.000	1.000	1.002
γ_{19}	1.001	1.000	1.000	1.000	1.003
γ_{29}	1.001	1.000	1.000	1.000	1.001

Table B.6: Potential scale reduction factor calculated for the β parameters for Models 6 – 10 fitted to the hoki catch weight data. It appears that the parameters β_{07} and $\beta 17$ for Model 7 have not converged to stationary distributions.

Parameter	Model 6	Model 7	Model 8	Model 9	Model 10
β_{01}	1.001	1.000	1.000	1.000	1.003
β_{11}	1.001	1.001	1.002	1.000	1.000
β_{02}	1.000	1.001	1.004	1.001	1.000
β_{12}	1.000	1.001	1.002	1.000	1.001
β_{03}	1.000	1.000	1.000	1.001	1.002
β_{13}	1.000	1.008	1.006	1.003	1.000
β_{04}	1.001	1.000	1.000	1.000	1.000
β_{14}	1.003	1.000	1.006	1.001	1.001
β_{05}	1.002	1.000	1.000	1.008	1.000
β_{15}	1.003	1.002	1.000	1.006	1.000
β_{06}	1.001	1.001	1.003	1.003	1.000
β_{16}	1.004	1.000	1.000	1.038	1.000
β_{07}	1.000	1.198	1.000	1.001	1.000
β_{17}	1.001	2.957	1.002	1.000	1.008
β_{08}	1.000	1.001	1.001	1.000	1.000
β_{18}	1.000	1.008	1.000	1.000	1.001
β_{09}	1.000	1.006	1.001	1.001	1.001
β_{19}	1.000	1.000	1.006	1.002	1.000

Table B.7: Potential scale reduction factor calculated for the γ parameters for Models 6 – 10 fitted to the hoki catch weight data. It appears that γ_{17} and γ_{27} for Model 7 have not converged to stationary distributions.

Parameter	Model 6	Model 7	Model 8	Model 9	Model 10
γ_{01}	1.000	1.006	1.000	1.002	1.000
γ_{11}	1.001	1.001	1.011	1.003	1.000
γ_{21}	1.007	1.006	1.005	1.001	1.000
γ_{02}	1.000	1.001	1.002	1.000	1.000
γ_{12}	1.001	1.001	1.001	1.000	1.008
γ_{22}	1.000	1.003	1.006	1.001	1.001
γ_{03}	1.002	1.009	1.000	1.000	1.000
γ_{13}	1.004	1.000	1.002	1.001	1.000
γ_{23}	1.000	1.004	1.005	1.003	1.000
γ_{04}	1.006	1.001	1.007	1.000	1.000
γ_{14}	1.014	1.003	1.006	1.003	1.001
γ_{24}	1.002	1.001	1.000	1.000	1.000
γ_{05}	1.000	1.000	1.000	1.011	1.004
γ_{15}	1.003	1.004	1.000	1.000	1.000
γ_{25}	1.001	1.002	1.000	1.006	1.001
γ_{06}	1.000	1.022	1.000	1.013	1.005
γ_{16}	1.002	1.019	1.000	1.046	1.003
γ_{26}	1.003	1.002	1.000	1.020	1.003
γ_{07}	1.009	1.004	1.004	1.003	1.001
γ_{17}	1.006	387.931	1.000	1.001	1.007
γ_{27}	1.000	70.092	1.000	1.001	1.015
γ_{08}	1.000	1.000	1.002	1.001	1.000
γ_{18}	1.003	1.008	1.000	1.002	1.000
γ_{28}	1.001	1.003	1.000	1.001	1.000
γ_{09}	1.003	1.000	1.000	1.000	1.000
γ_{19}	1.000	1.006	1.005	1.000	1.000
γ_{29}	1.000	1.002	1.018	1.002	1.001

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281).

Akima, H. (1978). A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software (TOMS)*, **4**(2), 148–159.

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm. .

Anderson, T. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)* (pp. 55–66).

Anderson, T. W. et al. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in probability and statistics*, , 1–24.

Anderson, T. W. et al. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, **1**(1), 135–141.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, **27**(2), 93–115.

Atkinson, P. M. & LLoyd, C. D. (2007). Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. *Computers & Geosciences*, **33**(10), 1285–1300.

Bagley, N. W., Ballara, S. L., O'Driscoll, R. L., Fu, D., & Lyon, W. S. (2013). *A Review of Hoki and Middle-depth Summer Trawl Surveys of the Sub-Antarctic, November December 1991-1993 and 2000-2009*. Ministry for Primary Industries Wellington, New Zealand.

Ball, G. & Hall, D. (1965). *ISODATA. A Novel Method of Data Analysis and*. Technical report, Pattern Classification. Technical Report, Stanford Research Institute, Menlo ....

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Banerjee, S. & Gelfand, A. E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *Journal of the American Statistical Association*, **101**(476), 1487–1501.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(4), 825–848.

Banerjee, S., Gelfand, A. E., & Sirmans, C. (2003a). Directional rates of change under spatial process models. *Journal of the American Statistical Association*, **98**(464), 946–954.

Banerjee, S., Wall, M. M., & Carlin, B. P. (2003b). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, **4**(1), 123–142.

Bel, L., Allard, D., Laurent, J., Cheddadi, R., & Bar-Hen, A. (2009). Cart algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Analysis*, **53**(8), 3082–3093.

Bernardo, J. M. & Smith, A. F. (1994). Bayesian theory. *John Willey and Sons. Valencia (España)*, .

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, , 192–236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20.

Biau, G. & Scornet, E. (2016). A random forest guided tour. *Test*, **25**(2), 197–227.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Bivand, R., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with r-inla with some extensions. *Journal of statistical software*, **63**, 1–31.

Blangiardo, M. & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. Chichester, United Kingdom: John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, **7**, 39–55.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(6), 493–507.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. Chapman and Hall/CRC.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, **7**(4), 434–455.

Bruno, F., Guttorp, P., Sampson, P. D., & Cocchi, D. (2009). A simple non-separable, non-stationary spatiotemporal model for ozone. *Environmental and ecological statistics*, **16**(4), 515–529.

Brunsdon, C., Fotheringham, A., & Charlton, M. (2002). Geographically weighted summary statistics—a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, **26**(6), 501–524.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, **28**(4), 281–298.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47**(3), 431–443.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Cameletti, M., Ignaccolo, R., & Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, **22**(8), 985–996.

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**(2), 109–131.

Caraux, G. & Gascuel, O. (1992). Bounds on distribution functions of order statistics for dependent variates. *Statistics & probability letters*, **14**(2), 103–105.

Ceci, M. & Appice, A. (2006). Spatial associative classification: propositional vs structural approach. *Journal of Intelligent Information Systems*, **27**(3), 191–213.

Chen, Y. (2012). On the four types of weight functions for spatial contiguity matrix. *Letters in Spatial and Resource Sciences*, **5**(2), 65–72.

Cocchi, D., Greco, F., & Trivisano, C. (2007). Hierarchical space-time modelling of pm10 pollution. *Atmospheric environment*, **41**(3), 532–542.

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical geology*, **20**(4), 405–421.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N. & Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**(448), 1330–1339.

Cressie, N. & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 209–226.

Cressie, N. & Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Cressie, N. A. (1993). *Statistics for spatial data*.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, **88**(11), 2783–2792.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**(514), 800–812.

De Cesare, L., Myers, D., & Posa, D. (2001). Estimating and modeling space–time correlation structures. *Statistics & Probability Letters*, **51**(1), 9–14.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, **26**(2), 403–413.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Ecker, M. D., De Oliveira, V., & Isakson, H. (2013). A note on a non-stationary point source spatial model. *Environmental and ecological statistics*, **20**(1), 59–67.

Ecker, M. D. & Oliveira, V. D. (2008). Bayesian spatial modeling of housing prices subject to a localized externality. *Communications in Statistics—Theory and Methods*, **37**(13), 2066–2078.

El-Harbawi, M. (2013). Air quality modelling, simulation, and computational methods: a review. *Environmental Reviews*, **21**(3), 149–179.

Elhorst, J. P., Fischer, M. M., & Getis, A. (2010). Handbook of applied spatial analysis. *Methods*, , 377–407.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fouedjio, F. (2017). Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, **31**(8), 1887–1906.

Fouedjio, F. (2018). A fully non-stationary linear coregionalization model for multivariate random fields. *Stochastic Environmental Research and Risk Assessment*, **32**(6), 1699–1721.

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental monitoring and assessment*, **189**(7), 316.

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, **86**(1), 1–28.

Francis, R. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research*, **18**(1), 59–71.

Furrer, R., Genton, M. G., & Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.

Gelfand, A. E. & Banerjee, S. (2017). Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and its Application*, **4**, 245–266.

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelfand, A. E. & Schliep, E. M. (2016). Spatial statistics and gaussian processes: A beautiful marriage. *Spatial Statistics*, **18**, 86–104.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, **24**(6), 997–1016.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2019). Geographical

random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, , 1–16.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, **97**(458), 590–600.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**(494), 746–762.

Gosoniu, L., Vounatsou, P., Sogoba, N., Maire, N., & Smith, T. (2009). Mapping malaria risk in west africa using a bayesian nonparametric non-stationary model. *Computational Statistics & Data Analysis*, **53**(9), 3358–3371.

Gramacy, R. (2016). lagp: large-scale spatial modeling via local approximate gaussian processes in r. *Journal of Statistical Software, Articles*, **72**(1), 1–46.

Gramacy, R. B. & Apley, D. W. (2015). local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, **24**(2), 561–578.

Gramacy, R. B., Niemi, J., & Weiss, R. M. (2014). Massively parallel approximate gaussian process regression. *SIAM/ASA Journal on Uncertainty Quantification*, **2**(1), 564–584.

Green, P. J. & Sibson, R. (1978). Computing dirichlet tessellations in the plane. *The computer journal*, **21**(2), 168–173.

Guttorp, P. & Schmidt, A. M. (2013). Covariance structure of spatial and spatiotemporal processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, **5**(4), 279–287.

Haralick, R. M. & Shapiro, L. G. (1985). Image segmentation techniques. In *Applications of Artificial Intelligence II*, volume 548 (pp. 2–10).: International Society for Optics and Photonics.

Heaton, M. J., Christensen, W. F., & Terres, M. A. (2017). Nonstationary gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics*, **59**(1), 93–101.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., de Jesus, J. M., Tamene, L., et al. (2015). Mapping soil properties of africa at 250 m resolution: Random forests significantly improve current predictions. *PloS one*, **10**(6).

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, **6**, e5518.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, , 382–401.

Hoff, P. D. & Niu, X. (2012). A covariance regression model. *Statistica Sinica*, , 729–753.

Hooten, M. B. & Hobbs, N. (2015). A guide to bayesian model selection for ecologists. *Ecological Monographs*, **85**(1), 3–28.

Hrafnkelsson, B. & Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics*, **10**(2), 179–200.

Huang, H.-C., Martinez, F., Mateu, J., & Montes, F. (2007). Model comparison and selection for stationary space–time models. *Computational Statistics & Data Analysis*, **51**(9), 4577–4596.

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and data engineering*, **16**(12), 1472–1485.

Hughes-Oliver, J. M., Gonzalez-Farias, G., Lu, J.-C., & Chen, D. (1998). Parametric nonstationary correlation models. *Statistics & probability letters*, **40**(3), 267–278.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, **31**(8), 651–666.

Johnson, R. & Wichern, D. (1992). Applied multivariate statistical analysis, new jersey: Prentice hall. *Simon e Schuster Company Upper Saddle River*, .

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis*, **2**(4), 719–733.

Jun, M., Katzfuss, M., Hu, J., & Johnson, V. (2014). Assessing fit in bayesian models for spatial processes. *Environmetrics*, **25**(8), 584–595.

Kalogirou, S. (2016). Destination choice of athenians: An application of geographically weighted versions of standard and zero inflated p oisson spatial interaction models. *Geographical Analysis*, **48**(2), 191–230.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, **90**(430), 773–795.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, **112**(517), 201–214.

Kim, H.-M., Mallick, B. K., & Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, **100**(470), 653–668.

Kodinariya, T. M. & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, **1**(6), 90–95.

Krige, D. G. (1951a). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, **52**(6), 119–139.

Krige, D. G. (1951b). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, **52**(6), 119–139.

Kuschel, G., Metcalfe, J., Winton, E., Guria, J., Hales, S., & K Rolfe, A. W. (2012). Updated health and air pollution in new zealand study: summary report. *Wellington: Health Research Council*, .

Lan, W., Fang, Z., Wang, H., & Tsai, C.-L. (2018). Covariance matrix estimation via network structure. *Journal of Business & Economic Statistics*, **36**(2), 359–369.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. "CRC" press.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**(6), 1659–1673.

LeSage, J. P. & Pace, R. K. (2001). Spatial dependence in data mining. In *Data mining for scientific and engineering applications* (pp. 439–460). Springer.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14.

Li, X. & Claramunt, C. (2006). A spatial entropy-based decision tree for classification of geographical information. *Transactions in GIS*, **10**(3), 451–467.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423–498.

Link, W. A. & Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in ecology and evolution*, **3**(1), 112–115.

Liu, J., Ma, Y., & Wang, H. (2020). Semiparametric model for covariance regression analysis. *Computational Statistics & Data Analysis*, **142**, 106815.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, **28**(2), 129–137.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., & Straif, K. (2013). The carcinogenicity of outdoor air pollution. *Lancet Oncology*, **14**(13), 1262.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1 (pp. 281–297).: Oakland, CA, USA.

Majumdar, A. & Gelfand, A. E. (2007). Multivariate spatial modeling for geostatistical data using convolved covariance functions. *Mathematical Geology*, **39**(2), 225–245.

Malerba, D., Appice, A., Varlaro, A., & Lanza, A. (2005a). Spatial clustering of structured objects. In *International Conference on Inductive Logic Programming* (pp. 227–245).: Springer.

Malerba, D., Ceci, M., & Appice, A. (2005b). Mining model trees from spatial data. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 169–180).: Springer.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, **67**, 68–83.

Marutho, D., Handaka, S. H., Wijaya, E., et al. (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication* (pp. 533–538).: IEEE.

Matérn, B. (1960). Spatial variation, volume 36 of. *Lecture Notes in Statistics*, .

Mateu, J. et al. (2015). *Spatial and spatio-temporal geostatistical modeling and kriging*, volume 998. John Wiley & Sons.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, **58**(8), 1246–1266.

McBratney, A., Hart, G., & McGarry, D. (1991). The use of region partitioning to improve the representation of geo statistically mapped soil attributes. *Journal of Soil Science*, **42**(3), 513–532.

Ministry for the Environment (2015). Particulate matter concentrations 2006–2013. Retrieved from `https://data.mfe.govt.nz/layer/2667-particulate-matter-concentrations-20062013/webservices/`.

Ministry for the Environment & Statistics New Zealand (2007). Environment new zealand 2007. Retrieved from `https://www.mfe.govt.nz/publications/environmental-reporting/environment-new-zealand-2007`.

Ministry for the Environment & Statistics New Zealand (2015). New zealand's environmental reporting series: Environment aotearoa

2015. Retrieved from http://www.mfe.govt.nz/publications/environmental-reporting/environment-aotearoa-2015.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**(1/2), 17–23.

Morris, L. (2017). Spatial and temporal modelling of hoki distribution using gaussian markov random fields. .

Mou, Y., He, Q., & Zhou, B. (2017). Detecting the spatially non-stationary relationships between housing price and its determinants in china: Guide for housing market sustainability. *Sustainability*, **9**(10), 1826.

Mukhopadhyay, S. & Sahu, S. K. (2018). A bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in england and wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**(2), 465–486.

NIMBLE Development Team (2017). *NIMBLE: An R package for programming with BUGS models, version 0.6-6.*

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., & Papritz, A. J. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, **4**(1), 1–22.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Citeseer.

Paciorek, C. J. et al. (2013). Spatial models for point and areal data using markov random fields on a fine grid. *Electronic Journal of Statistics*, **7**, 946–972.

Paciorek, C. J. & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, **17**(5), 483–506.

Parliamentary Library, NZ (2014). Obesity and diabetes in new zealand, accessed january 23, 2019. Retrieved from `https://www.parliament.nz/en/pb/research-papers/document/00PLLawRP2014041/obesity-and-diabetes-in-new-zealand`.

Pollice, A. (2011). Recent statistical issues in multivariate receptor models. *Environmetrics*, **22**(1), 35–41.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**(2), 181–199.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, , 1033–1048.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Requia, W. J., Coull, B. A., & Koutrakis, P. (2019). Evaluation of predictive capabilities of ordinary geostatistical interpolation, hybrid interpolation, and machine learning methods for estimating pm2. 5 constituents over space. *Environmental research*, **175**, 421–433.

Risser, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv preprint arXiv:1610.02447*, .

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

Rue, H. & Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of Statistical Planning and Inference*, **137**(10), 3177–3192.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2), 319–392.

Rychlik, T. (1992). Stochastically extremal distributions of order statistics for dependent samples. *Statistics & probability letters*, **13**(5), 337–341.

Sahu, S. K. & Bakar, K. S. (2012). Hierarchical bayesian autoregressive models for large space–time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry*, **28**(5), 395–415.

Sampson, P. D. (2014). Spatial covariance. *Wiley StatsRef: Statistics Reference Online*, .

Sampson, P. D. & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**(417), 108–119.

Schabenberger, O. & Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. CRC press.

Schmidt, A. M., Guttorp, P., & O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, **22**(4), 487–500.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, **406**, 109–120.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Scrucca, L. et al. (2005). Clustering multivariate spatial data based on local measures of spatial autocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica*, **20**(1), 11.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3), 485–493.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Stein, M. L. (2005). Statistical methods for regular monitoring data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(5), 667–687.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, **1**(804), 801.

Stojanova, D., Ceci, M., Appice, A., Malerba, D., & Džeroski, S. (2011). Global and local spatial autocorrelation in predictive clustering trees. In *International Conference on Discovery Science* (pp. 307–322).: Springer.

Szatrowski, T. H. (1980). Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *The Annals of Statistics*, , 802–810.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, **46**(sup1), 234–240.

US Environmental Protection Agency (2018). Air quality system data mart [internet database] accessed october 25, 2018. Retrieved

from https://www.epa.gov/outdoor-air-quality-data/download-daily-data.

Vehtari, A. & Gelman, A. (2014). Waic and cross-validation in stan. *Helsinki: Aalto University*, .

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, **27**(5), 1413–1432.

Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications.

Ward, E. J. (2008). A review and comparison of four commonly used bayesian and maximum likelihood model selection tools. *Ecological Modelling*, **211**(1-2), 1–10.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**(Dec), 3571–3594.

White, G. & Ghosh, S. K. (2009). A stochastic neighborhood conditional autoregressive model for spatial data. *Computational statistics & data analysis*, **53**(8), 3033–3046.

Willmott, C. J. & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, **30**(1), 79–82.

Yuan, Y. & Johnson, V. E. (2012). Goodness-of-fit diagnostics for bayesian hierarchical models. *Biometrics*, **68**(1), 156–164.

Yule, G. U. (1921). On the time-correlation problem, with especial reference to the variate-difference correlation method. *Journal of the Royal Statistical Society*, **84**(4), 497–537.

Zhang, P., Huang, Y., Shekhar, S., & Kumar, V. (2003). Exploiting spatial autocorrelation to efficiently process correlation-based similarity queries. In *International Symposium on Spatial and Temporal Databases* (pp. 449–468).: Springer.

Zou, T., Lan, W., Wang, H., & Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, **112**(517), 266–281.

Zwiernik, P., Uhler, C., & Richards, D. (2014). Maximum likelihood estimation for linear gaussian covariance models. *arXiv preprint arXiv:1408.5604*, .