Spatio-temporal modelling of localised health inequalities in Glasgow

Riham Hamza Ismail

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

2024

Abstract

The main aim of this thesis is to develop a statistical clustering methodology for disease mapping. Disease mapping studies aim to understand a disease's spatial pattern and identify areas with low or high disease risk. These studies play an essential role in epidemiology and public health by providing information on how disease exposures differ geographically and assisting in allocating resources for prevention or intervention strategies. Commonly, such studies are based on areal data, which partitions the study region into a set of non-overlapping sub-regions. The standard clustering techniques for grouping data ignore the spatial dependencies between nearby areas in areal data. Therefore, the first model proposed in this thesis incorporates the spatial information within a Poisson finite mixture model for clustering areal data. The disease data are usually available over multiple timepoints, providing a valuable opportunity to carry out examinations of temporal trends and patterns. Thus, the two other methods proposed in this thesis are both forms of spatio-temporal generalised additive mixed model designed to capture trends and variations over both temporal and spatial dimensions. The first of these approaches estimates the disease risk over time and then identifies the high and low-risk clusters of spatio-temporal disease risk data. The final model in this thesis considers the potential clustering structure in spatial data over time and thereafter estimates the disease risk. These models are each used to assess the spatial and temporal trends of COVID-19 cases in the Greater Glasgow and Clyde Health Board areas. A key finding was that areas in different clusters often exhibited similar temporal trends but somewhat different means. The study also clearly identified several waves

of COVID-19 cases during the study period, most notably an increase in COVID-19 cases in September 2021, potentially influenced by the UK's rules in managing the COVID-19 epidemic.

Declaration

I, Riham Hamza Ismail, declare that all the work presented in this thesis has been done by myself for the award of the degree of Doctor of Philosophy from the School of Mathematics and Statistics under the supervision of Dr Craig Anderson and Dr Nema Dean, except where otherwise stated. This thesis has not been submitted in part or complete to any other university or towards any other degree before.

Acknowledgements

I would like to express my deep appreciation to my supervisors, Dr Craig Anderson and Dr Nema Dean, for their guidance, support, kindness, and encouragement throughout the stage of my PhD. This work would not have been complete without them.

I would like to thank the Ministry of Higher Education and the University of Jeddah for funding my study. Also, I would like to thank the Saudi Arabian Cultural Bureau SACB for their support throughout my study period.

To my parent, Hamza and Maha, I can not thank you enough for all your love, constant support, and encouragement. I would like to thank my sisters, Raneem and Ranad, and all my relatives for their support and encouragement during my journey.

A special thanks to my friends, Bashayer, Hanadi, Shaykhah, and Shuhrah, for their support, help, and kindness. We shared together moments of deep anxiety, but we also had wonderful, unforgettable times. Your presence has added immeasurable value to my study and life in Glasgow.

A huge thanks to my husband, Salim; as I reflect on my journey, I realize I wouldn't be where I am today without your love, support, encouragement, and patience. Last but not least, I would like to thank my lovely children, Wateen and Yoseph, for their love, smile, and cheerfulness, which motivated me to complete this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Disease mapping is a statistical approach used to study spatial and spatio-temporal disease data and determine high and low disease risk across geographical regions (MacNab et al., 2006), which may indicate potential disease outbreaks. Grouping areas, in terms of their spatial variation in disease risk, has a significant impact on improving public health by helping the government and health officials to concentrate on better supporting people in high-risk areas with healthcare resources or enabling the researchers to pinpoint the causes of these health inequalities. These measures could manifest as a vaccination initiative or a public education drive regarding possible risk elements. The term health inequalities refers to the differences in disease risk across social and population groups (NHS-Scotland 2016). One of the most important reasons for these differences is socioeconomic, deprivation or poverty, where the low-risk areas are more likely to be affluent. In contrast, the high-risk areas include the most deprived areas. The other factors of these differences are usually related to environmental exposures (e.g. water quality, air pollution), population habits (e.g. smoking, exercise, diet), or physical geography (e.g. temperature).

One of the earliest research on disease mapping was by John Snow, who created the first map for the cholera epidemic in Soho, London, in 1854. In those days, although

people thought that the air spread this disease, Snow (1855) created a map (Figure 1.1) for the cases of cholera disease, which revealed that cases were found near a water pump. The map helped people determine that cholera is spread by contaminated water and focused on revamping sanitation and water supply systems.

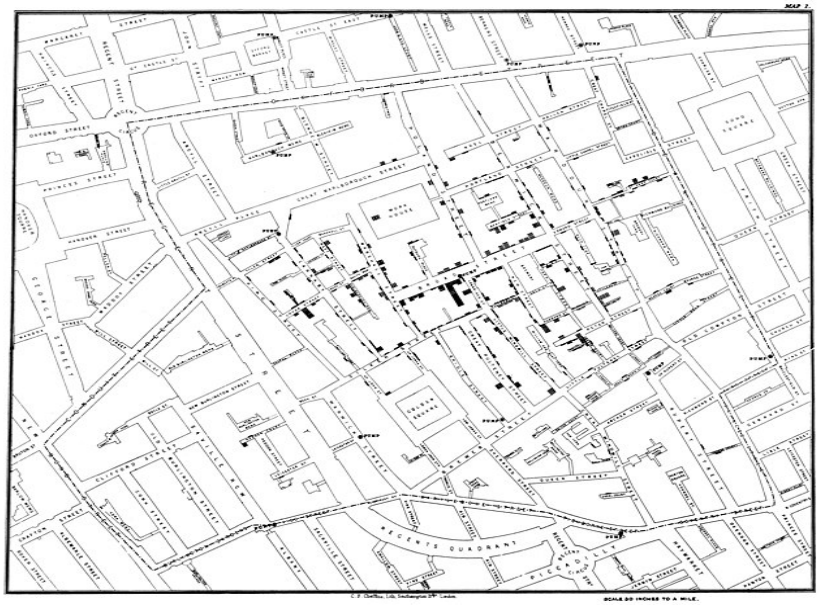


Figure 1.1: The map of cholera cases in Soho, London constructed by John Snow (Snow, 1855).

Commonly, disease mapping will be based on areal data, which partitions the study region into n non-overlapping sub-regions such as health board areas, and estimates the disease risks for each area. Usually, the total number of cases of a disease is available to the public instead of data at the personal level, in light of confidentiality reasons. Areas on the map will be in different colours or shades to represent the level of risk associated with diseases.

Conditional auto-regressive (CAR) models (Besag et al., 1991) are popular for estimating disease risk. The conditional auto-regressive (CAR) models assume spatial autocorrelation among adjacent areas where areas spatially close together can be more

similar than those distant ones as Tobler (1970) said *"The first law of geography: everything is related to everything else, but near things are more related than distant things"*. In this thesis, the correlation and spatial closeness between two areas are represented by a binary neighbourhood matrix $\boldsymbol{W}$, where $w_{ij} = 1$ if two areas are sharing a common border and $w_{ij} = 0$ if they do not share a common border. CAR models make the assumption of a consistent level of spatial autocorrelation across the study areas. However, this assumption, which forces areas close to each other to have similar risks, is not valid for all cases. Many researchers focus on developing models and approaches to deal with this matter.

One of the ideas is to partition areas into several disease risk clusters. Clustering is a set of techniques to allocate objects to groups, where the objects in each group are close to each other or share similar features and characteristics, and they are different from the objects in the other groups (Giordani, 2019). The goal of clustering is to find distinct groups in the data, but for spatial data, it may be better not to use standard clustering techniques since they will ignore the spatial dependencies between nearby regions. This thesis aims to develop a spatial clustering approach that can detect different areal clusters, considering their geographical information, and estimate disease risk. A spatio-temporal approach that can estimate the disease risk and identify disease risk clusters over a period of time will also be developed. The approaches proposed in this thesis are applied to Coronavirus disease (COVID-19) cases in the Greater Glasgow and Clyde Health Board region (Figure 1.2), which contains $n = 257$ administrative regions known as Intermediate Zones (IZs). COVID-19 data from the Scottish Health and Social Care Open Data are used throughout this thesis (https://www.opendata.nhs.scot/dataset/covid-19-in-scotland). Chapter 4 of this thesis focuses on one-time point, specifically data collected on June 10, 2020. This particular time point was selected because it coincided with the availability of COVID-19 test-

ing and heightened public awareness of the associated symptoms. Chapter 5 explores spatial-temporal data which includes a long period, 60 time points from August 2020 to October 2021. This long time frame allowed for a comprehensive analysis of the progression of COVID-19 over time and across different geographical regions. In Chapter 6, COVID-19 data during periods of fewer restrictions was of specific interest, where the proposed approach was applied to the period from March 2021 through October 2021.

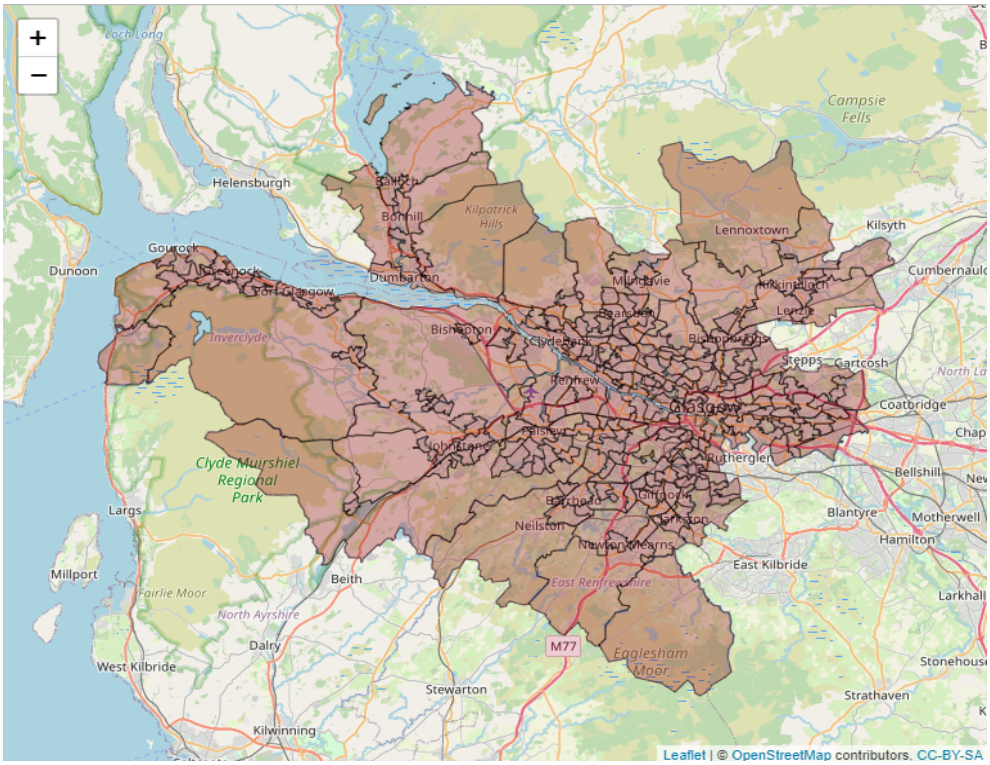


Figure 1.2: Map of the 257 Intermediate Zones (IZs) of the Greater Glasgow and Clyde Health Board.

Overview of the thesis

This thesis contains seven chapters. A brief preview of the contents of each chapter will be introduced here.

Chapter 2 outlines the general statistical methodology and inference methods which will be used across this thesis. Chapter 3 will introduce disease mapping and spatial and spatio-temporal literature along with a brief overview of the Coronavirus disease (COVID-19) epidemic.

In Chapter 4, a spatially constrained Poisson Finite Mixture model is developed and applied to simulated data and the Coronavirus disease cases in the Greater Glasgow and Clyde Health Board region, which will identify disease risk clusters taking into account the spatial information. This model will enable health authorities to identify high-risk areas at a specific time point and let them make some critical decisions before an outbreak. The spatial information will be incorporated in the Poisson Finite Mixture Model via a Gibbs prior, considering spatial dependencies between nearby areas (neighbours).

As disease risk data are usually available over a period of time, Chapter 5 will introduce a spatio-temporal generalised additive mixed model which fits spatially correlated random effects via the conditional autoregressive (CAR) model, while P-spline smoothing is used for the fixed and random temporal components to estimate the disease risk over time. Then, a Model-based clustering algorithm is used (Fraley & Raftery, 2002) to identify clusters from P-spline coefficients of the interaction between time and space. Detecting areas with increasing disease risk over time will enable public health officials to respond effectively to these trends and those areas at risk and reduce unnecessary ex-

penses in low-risk areas.

Considering the potential clustering structure in spatial data over time, Chapter 6 will present a two-stage approach which identifies the optimal number of clusters using the model-based clustering approach introduced in Section 2.11.1 and added the structure of the clusters as a factor to the spatio-temporal generalised additive mixed model, introduced in Chapter 5, to estimate the disease risk over time. Finally, Chapter 7 summarises the results of this thesis and discusses future research work. All the models were written in the R statistical language (R Core Team et al., 2013), utilizing version 4.0.0..

Chapter 2

Statistical background

This chapter outlines an overview of statistical theories and methodologies used in this thesis. Section 2.1 introduces frequentist statistics and inference methods for frequentist approaches. Section 2.2 presents Bayesian statistics with its inference methods and the idea of the prior and posterior distributions. A brief review of finite mixture models, generalised linear models, and generalised additive models are given in Section 2.3, 2.4, and 2.6. A review of spatial data and spatial modelling is outlined in Section 2.9. Section 2.10 introduces some common models in the field of spatio-temporal modelling. Finally, Section 2.11 will outline several clustering algorithms that have been used in this thesis.

2.1 Frequentist statistics

One of the most common inferential methods in statistics is the frequentist approach, which assumes that the model's parameters are unknown but can be estimated (Neyman, 1937). Under this approach, one can assume that a vector of independent observed data, $\mathbf{Y} = (Y_1, ..., Y_n)$, comes from a probability distribution $f(\mathbf{Y}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, which are unknown. These parameters can be estimated using the likelihood approach by maximising the likelihood function $L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^{n} f(Y_i|(\boldsymbol{\theta})$

(Zacks, 1981). In other words, the optimal estimated parameter value $\hat{\boldsymbol{\theta}}$ is the value globally or locally maximising the likelihood function. It is common and easier to use the log-likelihood function $l(\boldsymbol{\theta}|\mathbf{Y}) = \ln L(\boldsymbol{\theta}|\mathbf{Y})$ since we might need to find some derivative of the likelihood function when it is available, and the value that maximises the log-likelihood function will likewise maximise the likelihood function because the logarithm is a monotonically increasing function (Casella & Berger, 2002). The inference of $\boldsymbol{\theta}$ is based on a point estimate $\hat{\boldsymbol{\theta}}$ and calculating the c% confidence interval for the estimation uncertainty. These intervals are defined as c% of the intervals would contain the parameter's true value if the data were sampled repeatedly and intervals were created each time (Held & Sabanés Bové, 2014).

2.2 Bayesian Statistics

2.2.1 Introduction

In recent years, Bayesian statistics has become a more common approach to statistical inference. It is a branch of statistics where researchers can update their beliefs about random events once new data are available. Bayes' Theorem was developed in the 18th century by Thomas Bayes (Bayes, 1763), and is defined for two events A and B as follows:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}, \tag{2.2.1}$$

where $p(A)$ is the probability that the random event A occurs, $p(B)$ is the probability that the random event B occurs, $p(A|B)$ is the conditional probability of event A occurring given that event B has occurred, and $p(B|A)$ is the conditional probability of event B occurring given that event A has occurred. Bayes' theorem combines the likelihood of event B given event A and the prior distribution of event A to calculate a probability assessment regarding the distribution of event A given the occurrence of event B.

In Bayesian statistics, the parameters $\boldsymbol{\theta}$ are considered random variables where each parameter is assigned a distribution in advance ,prior distribution, $\pi(\boldsymbol{\theta})$. The inference about $\boldsymbol{\theta}$ is made in terms of probability statements conditional on the observed values $\mathbf{Y}$. This can be expressed as $p(\boldsymbol{\theta}|\mathbf{Y})$, which is known as a posterior distribution. This posterior distribution will summarise the information about the parameter θ.

The formula of Bayes' rule can be rewritten as follows:-

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{Y})}, \tag{2.2.2}$$

where $p(\mathbf{Y}|\boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior distribution. $p(\mathbf{Y})$ is a normalisation constant or the marginal distribution of the data, $p(\mathbf{Y}) = \int \pi(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ if $\boldsymbol{\theta}$ is continuous and $p(\mathbf{Y}) = \sum_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})$ if $\boldsymbol{\theta}$ is discrete. As $p(\mathbf{Y})$ does not depend on $\boldsymbol{\theta}$, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ can be obtained up to a constant of proportionality as follows:-

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\theta) \times \pi(\boldsymbol{\theta}), \tag{2.2.3}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution, which represents our prior knowledge about the parameter $\boldsymbol{\theta}$ (more detail in Section 2.2.2), and $p(\mathbf{Y}|\boldsymbol{\theta})$ is the data likelihood function.

The Bayesian approach has many advantages. It is flexible in modelling complex problems and dependencies between variables. It allows incorporation of prior knowledge or beliefs about the parameters being estimated. Moreover, Bayesian estimation provides a probability distribution for the parameter of interest conditional on the observed data (i.e., the posterior distribution) rather than just a point estimate. Obtaining posterior distribution parameters therefore allows for any function of the unknown parameters, and corresponding uncertainty, to be estimated.

The incorporation of prior information is an advantage of Bayesian inference, however it could be disadvantage where defining a prior distribution is difficult and/or different choices for a prior distribution have a strong influence on the posterior distribution. Another drawback of Bayesian inference often lies in the substantial computational time required, as posterior distributions without closed forms rely on Markov chain Monte Carlo approaches to estimate the posterior distribution. In Chapter 4 of this thesis, a fully Bayesian inference approach will be employed to estimate the model's parameters. However, in Chapters 5 and 6, parameter estimation will involve a combination of Bayesian and non-Bayesian approaches.

2.2.2 Prior Distribution

Section 2.2.1 introduces the idea of a prior distribution, where $\pi(\boldsymbol{\theta})$ displays the available information we know about the parameters $\boldsymbol{\theta}$ before observing $\mathbf{Y}$. As prior distributions have an essential role in determining the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$, it is crucial to select a suitable prior distribution to ensure reasonable parameters inference.

There are different types of prior distributions: informative, weakly informative or non-informative prior distributions. An informative prior is typically based on previous analyses, expert knowledge or existing literature. If we do not have prior knowledge about the parameter, a non-informative prior can be considered instead. For example, the uniform distribution on the interval [0,1] is used as a non-informative prior when used as a prior for a probability parameter since all possible values will be equally likely a priori. Alternatively, weakly informative prior distributions allow the researchers to use a little information rather than disregard all the information. A common weakly informative prior is a normal distribution with a large variance (e.g. $N \sim (0, 1000)$).

Also, Jeffreys (1946) introduced the Jeffreys prior, a form of weakly informative prior which is given by

$$\pi(\boldsymbol{\theta}) \propto \|I(\boldsymbol{\theta})\|^{\frac{1}{2}},$$

where $I(\boldsymbol{\theta})$ is the Fisher information which is defined as

$$I(\boldsymbol{\theta}) = E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{Y};\boldsymbol{\theta})\right)^2 \middle| \boldsymbol{\theta}\right].$$

If the prior and posterior distributions are from the same family, then the prior is described as being conjugate for that particular family or likelihood. Using a conjugate prior will lead to a closed-form expression which is computationally convenient. For example, consider a beta distribution as a prior $\pi(\theta) \sim Beta(\alpha,\beta)$ where the likelihood is $p(Y|\theta) \sim Bin(Y|n,\theta)$; using equation (2.2.3), the posterior will be

$$p(\theta|Y) \propto \theta^Y(1-\theta)^{n-Y} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}.$$

The posterior distribution is

$$p(\theta|Y) \sim Beta(Y+\alpha, n-Y+\beta), \tag{2.2.4}$$

which has the same family beta distribution as the prior, but different parameters.

2.2.3 Bayesian inference

Bayesian inference is based on combining the prior distribution, representing our beliefs, with observed data in order to find the posterior distribution and evaluate the model parameters. Inference about the model parameters from a posterior distribution can be easily done using a conjugate prior distribution as discussed in Section 2.2.2. Computing the posterior distribution will be straightforward and have a standard distribution form. For example, in equation (2.2.4) we have the closed form of the posterior distribution where $p(\theta|Y) \sim Beta(Y+\alpha, n-Y+\beta)$, so the posterior has mean $\frac{Y+1}{n+2}$, standard deviation $\sqrt{\frac{(Y+1)(n-Y+1)}{(n+2)^2(n+3)}}$ and mode $\frac{Y}{n}$.

On the other hand, computing the posterior distribution can be difficult in some cases where the distribution is not a standard form. In such cases, Markov chain Monte Carlo (MCMC) methods can be used to estimate the distributions of the model parameters by generating random samples of the parameters from a posterior distribution, where it creates a Markov chain which converges to a target posterior distribution after a sufficient number of iterations.

2.2.4 Metropolis-Hastings Algorithm and Gibbs Sampling (MCMC)

Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Hastings, 1970) is a random walk designed to converge to the specified target distribution. Consider the parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ and the data vector $\mathbf{y}$, the process of the algorithm is as follows:

1. Pick starting points $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.

For iteration $t = 1, ..., T$

2. Sample $\boldsymbol{\theta}^*$ from a proposal distribution $q_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$.

3. Calculate the acceptance probability

$$p = \min\left(\frac{p(\boldsymbol{\theta}^*|\mathbf{y})q_t(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})q_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}, 1\right).$$

4. Generate a uniform random sample

$$u \sim Uniform(0,1)$$

if $u \leq p$, accept the proposed $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t)}$ as $\boldsymbol{\theta}^*$;
otherwise, reject the proposed $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(t)}$ will be set equal to $\boldsymbol{\theta}^{(t-1)}$.

The Metropolis algorithm (Metropolis et al., 1953) is a special case of the Metropolis-Hastings algorithm with symmetric proposal distribution where $q_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = q_t(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)$, then the acceptance ratio will be written as follows:-

$$p = \min\left(\frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})}, 1\right).$$

For Metropolis-Hastings, choosing an appropriate proposal distribution that gives an optimal acceptance rate is essential. The acceptance rate will be high where the chance of the proposal being accepted is more likely, that is when the proposal distribution has a small variance, which will produce a value close to the current value. On the other hand, when the proposal distribution has a large variance, the acceptance rate

will be low because the proposed value is very different from the current value, and it will be unlikely to be accepted. Also, this cause problem with the mixing of the chain, where a high acceptance rate means the chain does not explore the full posterior density, and a low acceptance rate means the exploration is beyond the posterior density. In this thesis, the acceptance rate will be between 0.2 and 0.6 approximately (Roberts & Rosenthal, 1998), and to deal with the low and high acceptance rates, I will tune the proposal variance every 100^{th} iteration.

There are two common types of proposal distribution. The first type is the independent proposal distribution where each parameter is updated independently, meaning the proposal value θ^* does not depend on the current value $\theta^{(t-1)}$. The second common type is the random-walk proposal distribution, where the proposal value θ^* is generated from a distribution centred around the current value $\theta^{(t-1)}$.

Gibbs Sampling

In certain cases where we can sample directly from the conditional posterior distribution, one can use Gibbs sampling (Geman & Geman, 1984, Gelfand & Smith, 1990). Consider parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, where each parameter has full conditional distribution $(\theta_i|\theta_1, ..\theta_{i-1}, \theta_{i+1}, ..., \theta_p, \mathbf{y})$ with data vector $\mathbf{y}$, the steps of the Gibbs sampler are as follows:-

Set starting values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.

For iteration $t = 1, ..., T$,

1. sample $\theta_1^{(t)}$ from $p(\theta_1|\mathbf{y}, \theta_2^{(t-1)}, ..., \theta_p^{(t-1)})$

2. sample $\theta_2^{(t)}$ from $p(\theta_2|\mathbf{y}, \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)})$

...

p. sample $\theta_p^{(t)}$ from $p(\theta_p|\mathbf{y}, \theta_1^{(t)}, ..., \theta_{p-1}^{(t)})$

Repeat the step for T draws, where each draw represents the full set parameters values drawn from the conditional posterior distributions.

$$\begin{aligned}
\boldsymbol{\theta}^{(1)} &= (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_p^{(1)}) \\
\boldsymbol{\theta}^{(2)} &= (\theta_1^{(2)}, \theta_2^{(2)}, ..., \theta_p^{(2)}) \\
&. \\
&. \\
\boldsymbol{\theta}^{(T)} &= (\theta_1^{(T)}, \theta_2^{(T)}, ..., \theta_p^{(T)}).
\end{aligned}$$

Gibbs sampling is considered a spatial case of Metropolis-Hastings where the proposed state is accepted with a probability equal to 1. Given $p(\theta|y) = p(\theta_i, \theta_{-i}|y) = p(\theta_i|y, \theta_{-i})p(\theta_{-i}|y)$, then

$$\begin{aligned}
\alpha(\theta^*|\theta) &= \min\left\{1, \frac{p(\theta^* \mid y)p(\theta_i \mid y, \theta_{-i})}{p(\theta \mid y)p(\theta_i^* \mid y, \theta_{-i}^*)}\right\} \\
&= \min\left\{1, \frac{p(\theta_i^* \mid y, \theta_{-i}^*)p(\theta_{-i}^* \mid y)p(\theta_i \mid y, \theta_{-i})}{p(\theta_i \mid y, \theta_{-i})p(\theta_{-i} \mid y)p(\theta_i^* \mid y, \theta_{-i}^*)}\right\} \\
&= 1
\end{aligned}$$

similar terms will cancel each other and also $\theta_{-i}^* = \theta_{-i}$. The main benefit of Gibbs sampling is that proposals are always accepted. However, similar to other MCMC approaches, the primary drawback of Gibbs sampling is that the conditional probability distributions must be able to be derived.

Assessing Convergence

One should assess the convergence of chains, which means that the chains have approached a stationary distribution which approximates the true posterior distribution of the model parameters. Several visual and numerical diagnoses are used to check convergence, such as the trace plots and the Gelman-Rubin diagnostic (Gelman et al., 1992). Across this thesis, a trace plot will be used to determine the convergence. A trace plot is a plot of the parameter's value at each iteration against the iteration number, where the plot will look weakly stationary if there is no evidence of non-convergence. It is essential to discard the early iterations of a Markov Chain, known as a "burn-in" or "warm-up" period, to reduce the influence of the starting value because we want to estimate the parameters after we ensure that the chain has been converged. Figure 2.1 is an example of a trace plot for a parameter with 5,000 iterations. We can see convergence does not occur until after the first 500 iterations. This period is often known as burn-in, and is discarded from our analysis. It's worth mentioning that in Markov Chains, a correlation exists between consecutive draws from the posterior distribution. This correlation, known as autocorrelation, indicates that the samples are not independent. Thinning can be used to obtain independent samples. Thinning involves storing only every kth draw (after the burn-in period) from the posterior distribution while discarding the rest.

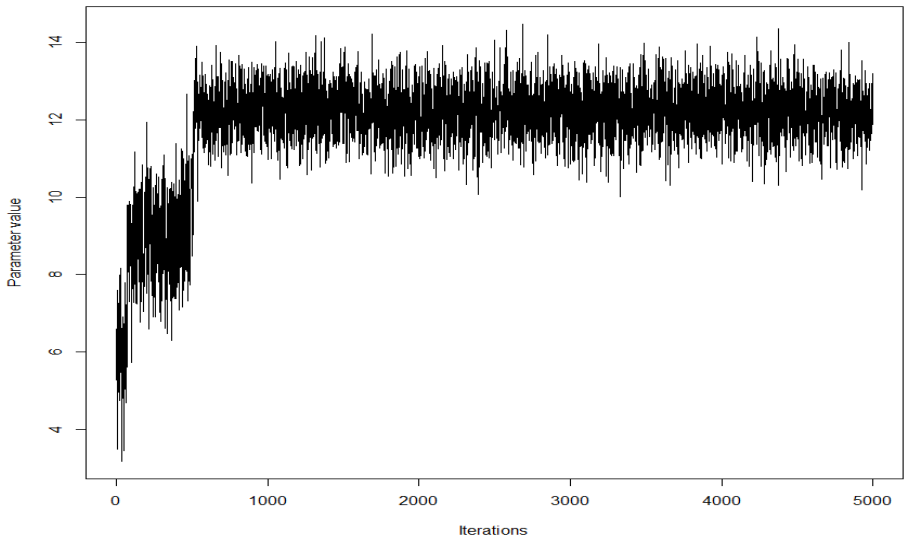


Figure 2.1: An example of a trace plot.

2.3 Finite Mixture Models

A finite mixture model (FMM) is a statistical model used in data analysis to represent a population as several groups or components, where data are generated from a distribution with different component densities according to a set of mixing proportions. One of the first uses of the finite mixture model was by the biometrician Karl Pearson (Pearson, 1894). Consider a random sample where $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)$ and $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ is the observed random sample where $\mathbf{y}_i = (y_{i1}, ..., y_{iP})$ is the observed values of the P-dimensional random vector $\mathbf{Y}_i$ which its density function is $f(\mathbf{y}_i)$. Then, the finite mixed model can be defined as follows:

$$f(\mathbf{y}_i) = \sum_{j=1}^{g} p_j \, f_j(\mathbf{y}_i) \qquad i = 1, \ldots, n \;\text{ and }\; j = 1, \ldots, g,$$

where g is the number of components, $\mathbf{p} = (p_1, ..., p_g)$ and p_j represent the probability

of being a member of the j^{th} component with

$$\sum_{j=1}^{g} p_j = 1,$$

and

$$0 < p_j < 1 \qquad j = 1, \ldots, g.$$

For example, the Poisson finite mixture model with g components is defined as follows:

$$f(\mathbf{y}_i \mid \mathbf{p}, \boldsymbol{\lambda}) = \sum_{j=1}^{g} p_j \frac{\lambda_j^{\mathbf{y}_i} e^{-\lambda_j}}{\mathbf{y}_i!} \qquad i = 1, \ldots, n \text{ and } \quad j = 1, \ldots, g.$$

In Chapter 4 and Chapter 6 of this thesis, the Poisson finite mixture model will be used.

2.4 Generalised linear model

The simplest regression model is the linear model, which estimates the linear relationship between the covariate data and the response variable. A standard linear model takes the form:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \mathbf{x}_i^\top \boldsymbol{\beta}, \end{aligned} \tag{2.4.1}$$

where Y_i is a response variable assumed to follow a Gaussian distribution with mean μ_i and variance σ^2, $\mathbf{x}_i^\top = (1, x_{i1}, ..., x_{iC})$ is a covariates vector containing C values of the covariates relating to the observation i and 1 represents the intercept term, $\boldsymbol{\beta}$ is a vector of the regression parameters which contain the slope coefficients of the covariates $(\beta_1, ..., \beta_C)$ and β_0 for the intercept term. However, the linear model is only appropriate when $\boldsymbol{Y}$ is assumed to be a Gaussian distribution. Where we have a non-normal (e.g., Poisson), we use generalised linear models (GLM), which were introduced by Nelder & Wedderburn (1972). The GLM is defined by identifying the response, which is a

member of the exponential family distribution, and the link function.

The exponential family

The distribution of Y_i with an assumed expectation μ_i belongs to the exponential family if it can be written in the following form:

$$f(y_i;\mu_i) = \exp[a(y_i)b(\mu_i) + c(\mu_i) + d(y_i)], \tag{2.4.2}$$

where a, b, c, d are known functions. Many distributions, such as Gaussian, Poisson, and Binomial, belong to this family.

Link function

The link function shows how the mean of the response $\mathbb{E}(Y_i) = \mu_i$ is related to a linear combination of the linear predictors $\mathbf{x}_i^\top \boldsymbol{\beta}$.

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{2.4.3}$$

$g(.)$ is a link function, which is a monotone and differentiable function. The choice of link function will depend on the distribution of the response data. For example, Gaussian data will take an identity link function where $g(\mu_i) = \mu_i$, and Poisson data use a natural log link function $g(\mu_i) = \ln(\mu_i)$.

Example

We can illustrate the GLM using an example of a Poisson distribution where the Poisson distribution is a member of the exponential family with a natural log link function. Let $Y_i \sim Poisson(\mu_i)$, the probability mass function is:

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!},$$

where $\mathbb{E}(Y_i) = \mu_i$. The Poisson distribution is a member of the exponential family since it can be written as:

$$f(y_i;\mu_i) = \exp[y_i \ln(\mu_i) - \mu_i - \ln(y_i!)]$$

which has the natural log link function. We can see that: $a(y_i) = y_i, b(\mu_i) = \ln(\mu_i), c(\mu_i) = -\mu_i$, and $d(y_i) = -\ln(y_i!)$.

2.5 Generalised Linear Mixed model

A generalised linear mixed model (GLMM) is a model that extends the generalised linear model (Section 2.4) by adding random effects along with the usual fixed effects. The random effects are used to account for correlation or non-independent observations. The model is defined as follows:

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \mathbf{b} \tag{2.5.1}$$

where $g(.)$ is a link function, which is a monotone and differentiable function, the mean of the response $\mu_i = \mathbb{E}(Y_i)$. $\mathbf{x}_i^\top \boldsymbol{\beta}$ represents the fixed effect term where $\mathbf{x}_i^\top = (1, x_{i1}, ..., x_{iC})$ is a covariates vector containing C values of the covariates relating to the observation i and 1 represents the intercept term, $\boldsymbol{\beta}$ is a vector of the regression parameters which contain the slope coefficients of the covariates $(\beta_1, ..., \beta_C)$ and β_0 for the intercept term. $\mathbf{Z}_i^\top \mathbf{b}$ represents the random effect term where $\mathbf{Z}_i^\top$ is a design vector and $\mathbf{b} = (b_1, ..., b_r)$ is a random effects vector which is assumed to have a normal distribution with mean zero and some unknown variance. Note that the linear mixed model (LMM) is a special case of a GLMM with an identity link function.

2.6 Generalised Additive Mixed Model

A generalised additive model (GAM) developed by (Hastie & Tibshirani, 1987) is a generalised linear model where the linear predictor includes one or more smooth functions of covariates. The model provides a very flexible specification of the relationship between the response and covariates. One can specify the model in terms of smooth functions instead of finding detailed parametric relationships. The general form of the model is as follows:

$$g(\mu_i) = \beta_0 + S_1(x_{1i}) + S_2(x_{2i}) + S_3(x_{3i}) + ... + S_m(x_{mi}) \tag{2.6.1}$$

Here,

$$\mu_i \equiv \mathbb{E}(y_i) \quad \text{and} \quad y_i \sim f(y_i; \mu_i),$$

where y_i is the response, β_0 is the intercept, and g(.) is a monotone and differentiable link function. The smooth function of a covariate x is represented by $S(.)$ (more details in section 2.7).

For the case of the mixed model, Lin & Zhang (1999) proposed generalised additive mixed modelling, which is defined as follows:

$$g(\mu_i) = \beta_0 + S_1(x_{1i}) + S_2(x_{2i}) + S_3(x_{3i}) + ... + S_m(x_{mi}) + \mathbf{Z}_i\mathbf{b} \tag{2.6.2}$$

where $\mathbf{Z}_i$ is a row of a random effects model matrix, $\mathbf{b}$ is a random effects vector.

Generalised additive mixed modelling is an extension of typical regression methods as it estimates the form of the relationship between a dependent variable and a number of given predictors. Instead of forcing the relation between a dependent variable and predictor to be linear, as is the case in typical linear regression, this relation is modelled

as a smooth function, which does not need to be linear.

2.7 Splines

In a generalised additive model, splines are often used to model the smooth function, which defines how the smooth terms vary over the predictor variables. There are several types of splines, such as cubic regression splines, basis splines (B-splines), and penalised basis splines (P-splines). In this thesis, P-splines have been used, penalising the second differences of the B-splines coefficients.

2.7.1 B-Splines

B-splines are polynomial functions joined together at points called knots (de Boor, Eilers & Marx, 1972, 1996). The knot sequence is a finite sequence of real numbers sorted in a non-decreasing order. Figure 2.2 illustrates B-spline curves fitted on a dataset for different numbers of knots, where we observe that increasing the number of knots will fit the data more and lead to a wiggly curve. In general, too many knots result in overfitting the data, while a few knots will lead to underfitting. There are different methods of choosing the number of knots, such as the Akaike information criterion (AIC) (Akaike, 1974).

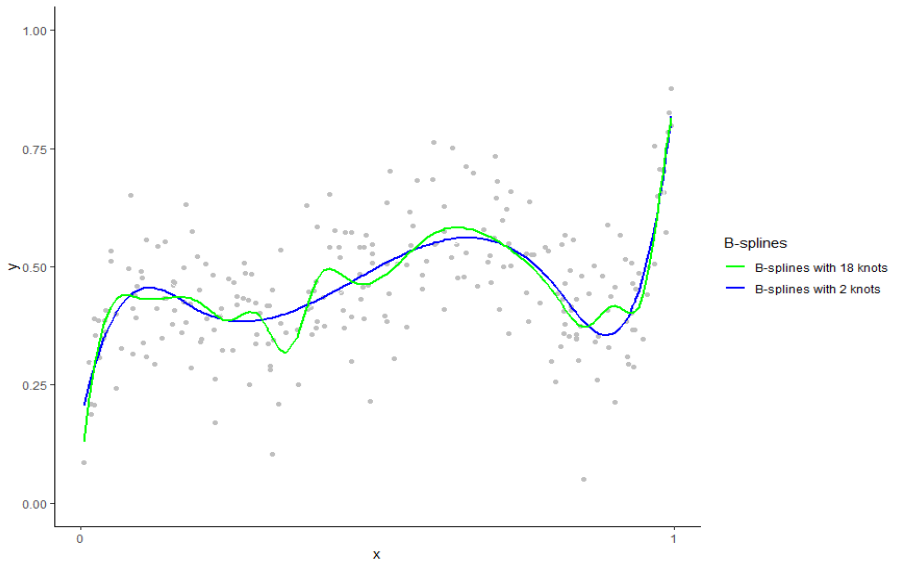


Figure 2.2: B-spline curves fitted on a simulated dataset for different numbers of knots. The blue curve represents a B-spline with 2 knots, whereas the green curve represents a B-spline with 18 knots.

Let $\Omega = (\omega_1, \omega_2, ..., \omega_{k+d+1})$ be a non-decreasing sequence where K is the number of the basis functions and d is the degree of the polynomial. The Ω represents the knots vector where ω_o, $o = 1, ..., K+d+1$ are the knots. A d degree B-Spline basis can be written as

$$S(x) = \sum_{k=1}^{K} p_k^d(x)\beta_k, \tag{2.7.1}$$

where p_k^d, k=1,..., K, are the B-spline basis functions, which are defined as:

$$p_k^{d+1}(x) = \frac{x-\omega_k}{\omega_{k+d}-\omega_k} p_k^d(x) + \frac{\omega_{k+d+1}-x}{\omega_{k+d+1}-\omega_{k+1}} p_{k+1}^d(x) \quad k = 1, .., K, \quad d > 0 \tag{2.7.2}$$

and

$$p_k^0(x) = \begin{cases} 1 & \omega_k \leq x < \omega_{k+1} \\ 0 & \text{otherwise.} \end{cases} \tag{2.7.3}$$

The splines are linear combinations of basis functions, so the spline coefficients will be estimated as estimating a linear model and can be written as follows:

$$S(x) = \sum_{k=1}^{K} p_k^d(x)\beta_k = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.7.4}$$

The B-Splines of the degree of d (Eilers & Marx, 1996) consists of $d+1$ polynomials of degree d. The derivatives are continuous up to $d-1$ at the internal knots. Furthermore, they are positive on a domain spanned by $d+2$ knots and zero everywhere else. Figure 2.3 shows the shapes of several types of B-spline bases, which are determined based on their degree. Degree 1 splines will have spike shapes, whereas degrees 2 and 3 will look like a smooth curve. Cubic splines create curves with continuous 2nd derivative, which makes it the most popular B-spline.

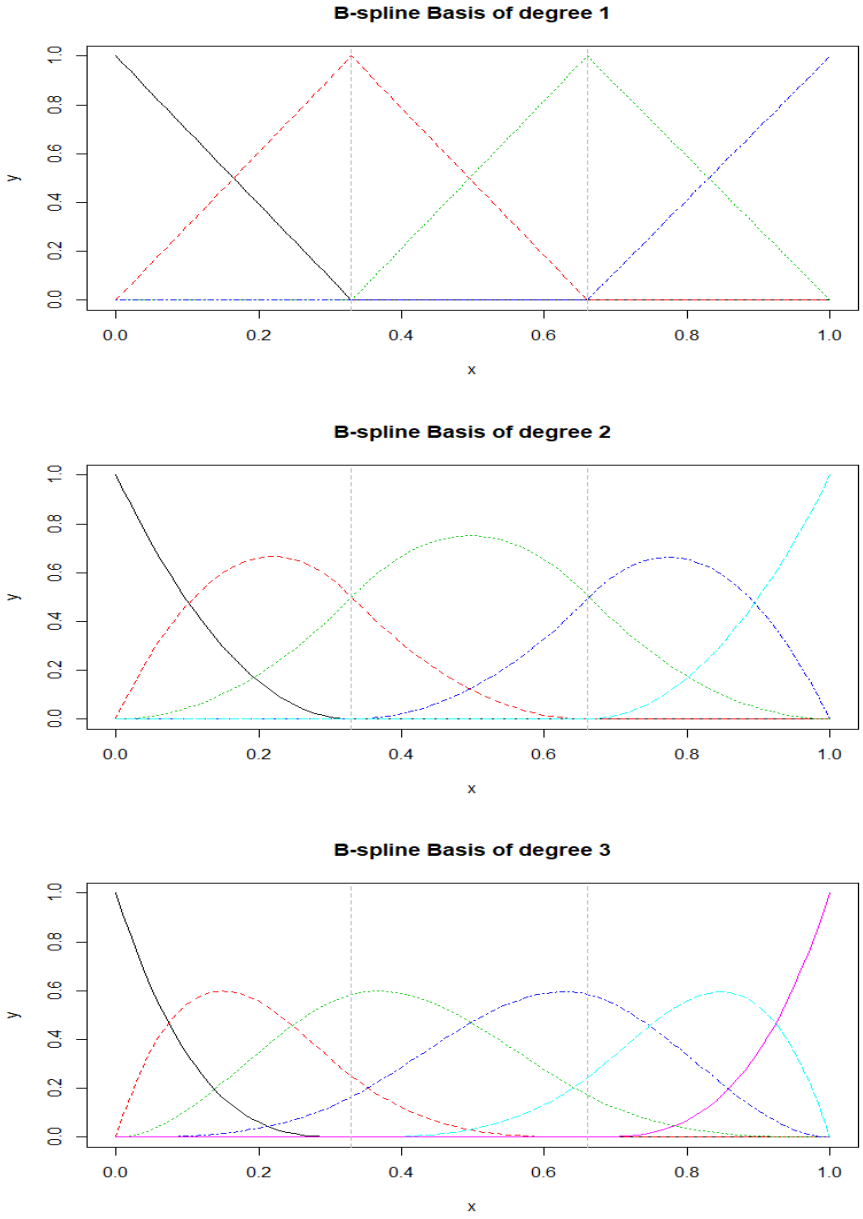


Figure 2.3: B-spline bases, from top to bottom: degree 1, degree 2, and degree 3 basis functions with two interior knots.

2.7.2 P-Splines

P-splines are penalised B-splines, which contain B-splines and a penalty to control "wiggliness", which means it balances the trade-off between the smoothness and over-fitting of the data. Let's say we have a regression of N data with K B-splines, the least squares objective function to minimize defined as

$$\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K} p_k(x_i)\beta_k\right]^2. \tag{2.7.5}$$

O'Sullivan(1986,1988) added a penalty on the second derivative of the curve

$$\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K} p_k(x_i)\beta_k\right]^2 + \gamma\int\left[\sum_{k=1}^{K} p_k^2(x_i)\beta_k\right]^2 dx. \tag{2.7.6}$$

where the parameter γ controls the smoothness such that when the value of γ increases, the smoother the results will be. Eilers & Marx (1996) proposed a penalty on higher-order of differences between adjacent β_k, so the objective function will be defined as follows:

$$\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K} p_k(x_i)\beta_k\right]^2 + \gamma\sum_{k=o+1}^{K}(\Delta^o\beta_k)^2. \tag{2.7.7}$$

where $\Delta^o\beta_k = \Delta(\Delta^{o-1}\beta_k)$ is the o^{th} order differences. In general, $\boldsymbol{D}_o\boldsymbol{\beta} = \Delta^o\boldsymbol{\beta}$, $\boldsymbol{D}_o$ is a $(K-o)\times K$ matrix and $\boldsymbol{\beta}$ is a vector of spline coefficients. The first order difference is calculated as $\Delta^1\beta_k = \beta_k - \beta_{k-1}$ with a difference matrix

$$D_1 = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & -1 & 1 & 0 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \tag{2.7.8}$$

We calculate the higher-order differences by applying the first-order differences repeatedly. For instance, calculating the second-order differences and their matrix will be as follows:

$$\begin{aligned}\Delta^2\beta_k &= \Delta(\Delta\beta_k)\\ &= (\beta_k - \beta_{k-1}) - (\beta_{k-1} - \beta_{k-2})\\ &= \beta_k - 2\beta_{k-1} + \beta_{k-2}\end{aligned} \tag{2.7.9}$$

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & \dots & 1 & -2 & 1 & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 \end{bmatrix} \tag{2.7.10}$$

We can rewrite the equation (2.7.7) as follows:

$$S = \|\mathbf{y} - \mathbf{p}\boldsymbol{\beta}\|^2 + \gamma\|\boldsymbol{D_o}\boldsymbol{\beta}\|^2 \tag{2.7.11}$$

$$S = (\mathbf{y} - \mathbf{p}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{p}\boldsymbol{\beta}) + \gamma\boldsymbol{\beta}^\top\boldsymbol{D_o^\top}\boldsymbol{D}_o\boldsymbol{\beta}. \tag{2.7.12}$$

After taking the first derivative of the equation (2.7.12) with respect to $\boldsymbol{\beta}$ and setting it equal to zero we obtain

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{p}^\top\boldsymbol{p} + \gamma\boldsymbol{D_o}^\top\boldsymbol{D_o}\right)^{-1}\boldsymbol{p}^\top\mathbf{y} \tag{2.7.13}$$

For the first-order differences $o = 1$ the penalty term in equation (2.7.7) would be

$$\gamma \sum_{k=2}^{K} (\beta_k - \beta_{k-1})^2$$

and the respective minimisation problem will be given by

$$\min_{\beta} \sum_{i=1}^{N} \left[y_i - \sum_{k=1}^{K} p_k(x_i)\beta_k \right]^2 + \gamma \sum_{k=2}^{K} (\beta_k - \beta_{k-1})^2. \tag{2.7.14}$$

The penalty term in equation (2.7.14) can also be stated in matrix notation as in equation (2.7.13) where the squares of differences of the first order difference is:

$$D_1^\top D_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots \\ -1 & 2 & -1 & 0 & \dots \\ 0 & -1 & 2 & -1 & \dots \\ 0 & 0 & -1 & 2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

For the second-order differences $o = 2$, the the penalty term in equation (2.7.7) is defined as:

$$\gamma \sum_{k=3}^{K} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2$$

The resulting minimisation problem is defined as:

$$\min_{\beta} \sum_{i=1}^{N} \left[y_i - \sum_{k=1}^{K} p_k(x_i)\beta_k \right]^2 + \gamma \sum_{k=3}^{K} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2. \tag{2.7.15}$$

For equation (2.7.13), the squares of differences of the second-order in matrix notation is defined as follows:

$$D_2^\top D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \dots \\ -2 & 5 & -4 & 1 & 0 & 0 & \dots \\ 1 & -4 & 6 & -4 & 1 & 0 & \dots \\ 0 & 1 & -4 & 6 & -4 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The parameter γ controls the smoothness where γ with a high value will cause over-smoothing, whereas a very low value of γ will lead to the fitted curve being too wiggly (Figure 2.4). In such cases, one can use some useful criteria to find an optimal value of the smoothing parameter, such as the Akaike information criterion (AIC), (generalized) cross-validation (GCV), or restricted maximum likelihood (REML) parameter estimation.

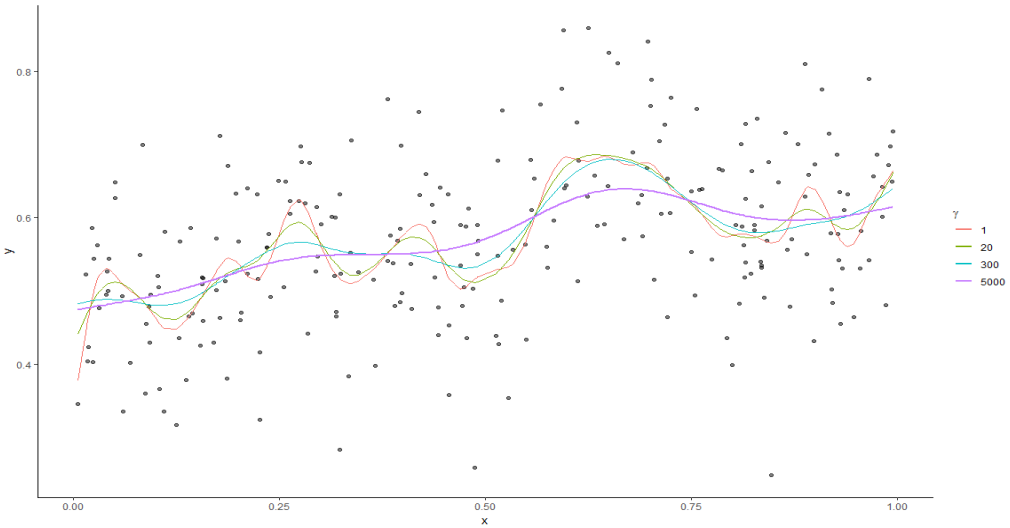


Figure 2.4: P-splines with different values of the smoothing parameter γ of a simulated dataset.

Choosing the smoothing parameters using restricted maximum likelihood (REML)

Consider a simple smoothing model

$$\boldsymbol{y} = f(\boldsymbol{x}) + \varepsilon, \quad f(\boldsymbol{x}) = \mathbf{p}\boldsymbol{\beta}$$

and re-parameterise

$$\boldsymbol{y} = \mathbf{p}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$$\boldsymbol{\beta} = \boldsymbol{T} \begin{bmatrix} \theta \\ a \end{bmatrix}$$

where θ corresponds to the part of the smooth function not penalised by the penalty matrix $\boldsymbol{P} = \boldsymbol{D}_o^\top \boldsymbol{D}_o$, $\boldsymbol{D}_o$ is a difference matrix with order o, and a is orthogonal to θ and it is penalised by the penalty matrix $\boldsymbol{P}$. The Singular Value Decomposition (SVD) of the penalty has been used to construct it:

$$\boldsymbol{T} \begin{bmatrix} \boldsymbol{U}_n & : & \boldsymbol{U}_s \end{bmatrix} \Rightarrow \theta = \boldsymbol{U}_n' \boldsymbol{\beta} \quad a = \boldsymbol{U}_s' \boldsymbol{\beta}$$

$$\boldsymbol{U}_n' \boldsymbol{P} \boldsymbol{U}_n = \mathbf{0} \Rightarrow \boldsymbol{\beta}' \boldsymbol{P} \boldsymbol{\beta} = \boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a}$$

$$\mathbf{p}\boldsymbol{\beta} = \mathbf{p}\boldsymbol{T} \begin{bmatrix} \theta \\ a \end{bmatrix} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{a}$$

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{Z}\boldsymbol{a})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{Z}\boldsymbol{a}) + \gamma \boldsymbol{a}^\top \Sigma \boldsymbol{a}. \tag{2.7.16}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{a} + \varepsilon, \;\; a \sim N(\mathbf{0}, \sigma_a^2 \Sigma), \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$$\gamma = \frac{\sigma^2}{\sigma_a^2}$$

Estimates of θ and a follow standard mixed model theory:

$$\hat{\theta} = (\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V}^{-1}\boldsymbol{y}$$

$$\hat{a} = \boldsymbol{GZ'V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}), \quad \boldsymbol{V} = \sigma^2\boldsymbol{I} + \boldsymbol{ZGZ'}$$

where $\boldsymbol{G}$ is the random effects covariance matrix. The smoothing parameters are selected by maximizing the restricted log likelihood (REML),

$$-\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X'V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V}^{-1})\boldsymbol{y}.$$

2.8 Model Comparison

For evaluating and comparing multiple statistical models for the same data, several methods have been proposed. Relying on maximising the likelihood of model selection could be problematic since adding more parameters tends to increase the likelihood, which may lead to overfitting. Hence, one can consider the information criteria that aim to balance maximising the likelihood and overfitting, such as the Akaike information criterion (AIC) (Akaike, 1974). Also, in this thesis, we will check the performance of models using root mean square error (RMSE) (Section 2.8.4).

2.8.1 Akaike Information Criterion

The Akaike information criterion (AIC) (Akaike, 1974) is a statistical measure used for model selection based on balancing the ability of a model to explain the data and the number of parameters. It is computed as follows:

$$AIC = -2\ln(\hat{L}) + 2K, \tag{2.8.1}$$

where $\hat{L}$ is the maximum value of the model likelihood, and K is the number of independent model parameters. It is preferable to choose the lowest AIC when comparing multiple models.

2.8.2 Bayesian Information Criterion

Another model selection criterion is the Bayesian information criterion (BIC) (Schwarz, 1978), which is expressed as follows:

$$BIC = -2\ln(\hat{L}) + K\ln(n) \tag{2.8.2}$$

where $\hat{L}$ is the maximum likelihood value, K is the number of independent model parameters, and n is the number of data points. The difference between the AIC and BIC is that the latter penalises the number of parameters more strongly, where adding an extra unnecessary parameter will lead to an increase in the second part of BIC, which has $\ln(n)$. We often prefer the model with the lowest BIC when we compare several models.

2.8.3 Deviance Information Criterion

The Deviance Information Criterion (Spiegelhalter et al., 2002) is another comparison method derived from model deviance and mostly used with Bayesian models. The DIC is defined as

$$DIC = \bar{D} + p_d, \tag{2.8.3}$$

where $\bar{D} = E[-2\ln(\hat{L})]$, which is the expectation of the posterior deviance and p_d is the effective number of parameters. Again, similar to AIC and BIC, the lowest DIC is preferred.

2.8.4 Root mean square error (RMSE)

Root mean square error (RMSE) is a statistical metric widely used to check model accuracy, where it calculates the average magnitude of the difference between the estimated and the true values. Assume n data points, the formula of the RMSE will be defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.8.4}$$

where y_i is the true value of the i^{th} data point and $\hat{y}_i$ is the estimated value of the i^{th} data point. A low value of RMSE indicates more accuracy.

2.9 Spatial modelling

Spatial data are any data that are linked to a geographical location. There are three main types of spatial data: geostatistical processes, point processes, and areal processes. Geostatistical processes are data that can be observed at many precise locations, such as air pollution, where it could be observed in several monitoring stations. Point processes apply where the locations themselves are the data, for example, the location of trees in a forest. Areal processes (areal data) occur when the study region is divided into non-overlapping sub-locations such as administrative zones, electoral wards, or council regions. For instance, my data in this thesis is the number of COVID-19 cases observed in 257 sub-regions in the Greater Glasgow and Clyde Health Board region,

which provides an aggregated summary of cases in each area unit. Since areal data keep the patient anonymous, it has been commonly used in health applications. In this thesis, I will focus on areal data modelling.

2.9.1 Spatial data

Suppose we have areal unit data within a study region **B**, which has been partitioned into a set of n non-overlapping sub-regions such that $\mathbf{B} = (B_1, ., B_n)$. The data are observed for each sub-region $\mathbf{y} = (y(B_1), ..., y(B_n))$ which will be denoted as $\mathbf{y} = (y_1, ..., y_n)$. Given that data are collected over space, it is reasonable to assume the existence of spatial correlations between close areas. An extension of the Poisson log-linear model (2.4) is commonly used to model the areal data and consider the spatial pattern.

2.9.2 Model

A hierarchical model for the areal unit using a Poisson log-linear model is given by

$$\begin{aligned} y_i &\sim \text{Poisson}(\mu_i) \ \text{ for } \ i = 1, ..., n \\ \log(\mu_i) &= \mathbf{x_i}^\top \boldsymbol{\beta} + \phi_i, \end{aligned}$$

where the explanatory variables for areal unit i are denoted by $\mathbf{x_i}$, and $\boldsymbol{\beta}$ is a vector of the coefficients. The random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ account for the spatial autocorrelation in the data, where ϕ_i is the random effect for areal unit i. These random effects are modelled using a conditional auto-regressive (CAR) prior distribution. CAR models control for spatial correlation via a $n \times n$ symmetric neighbourhood matrix **W**, which is specified as $w_{ij} = 1$ if the two areas i and j share a common border and $w_{ij} = 0$ if they do not share a common border. Several CAR prior distributions have been introduced,

such as the Intrinsic conditional autoregressive (ICAR) model.

Intrinsic conditional autoregressive

The intrinsic conditional autoregressive (ICAR) is the simplest CAR prior proposed by Besag et al. (1991). Under this model, the full conditional distribution for ϕ_i is defined as follows:

$$\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{W} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right) \tag{2.9.1}$$

where τ^2 controls the conditional variation between the random effects, $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, .., \phi_n)$ and the conditional expectation of ϕ_i is the mean of the random effect in neighbouring areal units, so each area will be modelled as similar as its neighbours. The ICAR distribution is an improper multivariate Gaussian distribution and does not take the strength of the correlation into account, which means it is only suitable for data with strong spatial autocorrelation rather than data with weak autocorrelation (Lee, 2011). The conditional distributions of $\boldsymbol{\phi}$ correspond to a multivariate Gaussian distribution that takes the following form:

$$\boldsymbol{\phi} \sim N\left(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W})^{-1}\right) \tag{2.9.2}$$

where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$ and $\mathbf{W1}$ is a vector containing the number of neighbours for each areal units. Several alternative CAR models were proposed to overcome this problem with the ICAR model, such as the convolution, Cressie, and Leroux CAR models.

Convolution CAR model

The convolution model or (BYM) CAR model proposed by Besag et al. (1991) is based on a linear combination of a spatially correlated random effect (ICAR) model with a set of independent random effects, and it is defined as:

$$\phi_i = \phi_{1,i} + \phi_{2,i}$$

$$\phi_{2,i} \sim N(0, \tau_2^2)$$

$$\phi_{1,i} | \boldsymbol{\phi}_{-1,i}, \boldsymbol{W} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_{1,j}}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau_1^2}{\sum_{j=1}^{n} w_{ij}} \right) \tag{2.9.3}$$

where $\boldsymbol{\phi}_1 = (\phi_{1,1}, ..., \phi_{1,n})$ is a set of random effects which follow the intrinsic CAR model (2.9.2) and $\boldsymbol{\phi}_2 = (\phi_{2,1}, ..., \phi_{2,n})$ is an independent and identically normally distributed random effect with zero mean and τ_2^2 variance. The ratio of the variances $\frac{\tau_1^2}{\tau_2^2}$ defines the level of the strength of the spatial autocorrelation between the random effects. Nevertheless, in the convolution model, one has to estimate two random effects $(\phi_{1,i}, \phi_{2,i})$ for each data point, but only their sum $\phi_{1,i} + \phi_{2,i}$ can be estimated from the data.

Stern and Cressie CAR model

Stern & Cressie (2000) proposed a proper CAR model adapted from the ICAR model (2.9.2) with an additional parameter to control the level of the spatial autocorrelation between the random effects donated by ρ. The model takes the form

$$\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{W} \sim N\left(\rho\frac{\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right) \tag{2.9.4}$$

when $\rho = 0$ means that we have independent random effects, whereas $\rho = 1$ corresponds to a strong spatial correlation and will be similar to the ICAR.

Leroux CAR model

The Leroux CAR model presented by Leroux et al. (2000) is given by

$$\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{W} \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}\right) \tag{2.9.5}$$

$\rho \in [0,1]$ controls the level of the spatial autocorrelation between the random effects in the same way as the Stern and Cressie model (2.9.4). If ρ is close to 1, the model will indicate a strong spatial autocorrelation between the random effects. On the other hand, if ρ is close to 0, it means lower spatial autocorrelation between the adjacent areas. The Leroux model (2.9.5) theoretically appeals more than the Stern and Cressie model (2.9.4) because when $\rho = 0$, the conditional variance of ϕ_i is equal to τ^2, so there is no additional spatial information about ϕ_i. The parameter ρ may assigned a hyper-prior, which leads to faster MCMC inference than specifying a continuous distribution (Lee, 2011).

2.10 Spatio-temporal modelling

Previously, in Section 2.9, we introduced spatial modelling where the data collected from n non-overlapping sub-locations at a single time point. However, disease risk data are usually available over a period of time. Spatio-temporal modelling, a statistical approach that captures spatial and temporal patterns, has become a topic of interest for researchers. One of the goals of these approaches is to identify the different temporal trends in the response across the study area. In this Section, Spatio-temporal modelling will be briefly introduced and more details about Spatio-temporal modelling in disease mapping are given in Section 3.4.

2.10.1 Bernardinelli model

One of the earliest models to address space and time is Bernardinelli's model (Bernardinelli et al., 1995). A Poisson generalised linear model was used in this paper and enabled a space and time interaction where varying temporal trends exist in different areas. Let $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)$ where $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iT})$ is the response for areal unit i for a period of time T. The model takes the following form:

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\mu_{it}) \text{ for } i = 1, ..., n, \ t = 1, ..., T, \\ \ln(\mu_{it}) &= \nu + \phi_i + (\beta + \delta_i)t, \end{aligned}$$

where ν is a global intercept term common for all areal units, ϕ_i is the spatial random effect of area i, β is the overall time effect, and δ_i is the term for space and time interaction. $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$ can be either independent (unstructured) random effects drawn from Gaussian distribution $N(0, \sigma^2)$ or structured random effects modelled using the intrin-

sic CAR prior introduced in (Section 2.9.2).

2.10.2 MacNab and Dean model

MacNab & Dean (2001) proposed a generalised additive mixed model combining the conditional autoregressive (CAR) model outlined in (Section 2.9) and B-splines outlined in (Section 2.7.1) for investigating spatial pattern and smoothing temporal trend, respectively, to estimate the disease risk. The model has the flexibility to capture complex temporal trends using B-splines. Consider a study region partitioned into n subregions and data collected for T time points where Y_{it} is the observed number of cases in area i at time t. Then, the model will take the following form:

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\mu_{it}) \text{ for } i = 1,...,n, \ t = 1,...,T, \\ \ln(\mu_{it}) &= \nu + \phi_i + S_0(t) + S_i(t), \end{aligned}$$

where ν is the overall intercept, ϕ_i is the spatial random effect for area i, $S_0(t)$ is a fixed temporal effect for all areas, and $S_i(t)$ is an area specific temporal effect. The CAR prior was used to model the spatial effect and B-splines for the global temporal trend $S_0(t)$. For the spatio-temporal interaction term MacNab & Dean (2001) considered two approaches. The first approach models $S_i(t)$ as a linear temporal trend using $S_i(t) = S_i t$. The second approach models $S_i(t)$ using B-splines for each area. Using a linear temporal trend is simpler than using B-splines to model $S_i(t)$ since B-splines require many parameters to be estimated; however, when considering a long period, using B-splines will likely provide a better fit to data.

2.11 Clustering algorithms

Clustering is a branch of statistics involving allocating objects to groups where the objects in each group share similar features and characteristics and differ from the objects in the other groups. For example, cluster analysis in disease mapping helps group areas with similar disease risk levels together, allowing us to identify the high-risk group. Several techniques for clustering are widely used in research, such as model-based and hierarchical clustering. In this Section, we will review some clustering techniques that have been used across this thesis.

2.11.1 Model-based clustering

Model-based clustering (Fraley & Raftery, 2002) is based on finite mixtures of distributions, where each component mixture density corresponds to a different cluster. Given data $\mathbf{y} = (\boldsymbol{y_1}, ..., \boldsymbol{y_n})$, the general form of the finite mixture of distributions with g groups is defined by

$$f(\mathbf{y}_i) = \sum_{j=1}^{g} \pi_j f_j(\mathbf{y}_i), \tag{2.11.1}$$

where π_j is the mixing proportion with $\sum_{j=1}^{g} \pi_j = 1$ and $0 < \pi_j \leq 1$, and $f_j(.)$ is the probability density function for the j^{th} group. Often, the mixture density components belong to the same parametric family, so the formula can be written as follows:

$$f(\mathbf{y}_i) = \sum_{j=1}^{g} \pi_j f(\mathbf{y}_i|\boldsymbol{\theta}_j) \tag{2.11.2}$$

where $\boldsymbol{\theta}_j$ is the vector of the parameters for the j^{th} group. A Gaussian mixture model is frequently used for continuous data, which takes the following form:

$$f(\mathbf{y}_i) = \sum_{j=1}^{g} \pi_j \phi(\mathbf{y}_i|\mu_j, \Sigma_j) \tag{2.11.3}$$

where π_j is is the prior probability of membership of group j and $\phi(\mathbf{y}_i|\mu_j,\Sigma_j)$ is the density of a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j.

In model-based clustering for frequentist inference, the expectation maximization (EM) algorithm (Dempster et al., 1977) is often used for estimating mixture model parameters, and the Bayesian Information Criterion (BIC) (Schwarz, 1978) is often used to select the optimal number of clusters or components (Section 2.8.2).

Expectation Maximization algorithm

The algorithm has two steps: The expectation step, which is called the E-Step and the maximization step, which is called the M-Step. First we define latent variables $\mathbf{z}$ to complete the observed data vector $\mathbf{y}$ where

$$z_{ij} = \begin{cases} 1 & \text{if } y_i \text{ belongs to component } j, \\ 0 & \text{otherwise.} \end{cases}$$

then the log-likelihood of the complete data will be defined as

$$\log L_c(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{z}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}(\log \pi_j + \log f_j(\mathbf{y_i}|\boldsymbol{\theta}_j)) \tag{2.11.4}$$

The E-Step computes the conditional expectation $\hat{z}_{ij}$ based on the current parameters value as follows:

$$\hat{z}_{ij}^{(t)} = \frac{\hat{\boldsymbol{\pi}}_j^{(t-1)} f_j(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_j^{(t-1)})}{\sum_{l=1}^{g} \hat{\boldsymbol{\pi}}_l^{(t-1)} f_l(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_l^{(t-1)})}$$

The M-Step maximizes the complete log-likelihood (2.11.4) with respect to the model parameters. If we assume $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then we can estimate the parameters as follows:

$$n_j^{(t)} = \sum_{i=1}^{n} \hat{z}_{ij}^{(t)}$$

$$\hat{\pi}_j^{(t)} = \frac{n_j^{(t)}}{n}$$

$$\hat{\mu}_j^{(t)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(t)} \boldsymbol{y}_i}{n_j^{(t)}}$$

$$\hat{\Sigma}_j^{(t)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(t)} \left(\boldsymbol{y}_i - \hat{\mu}_j^{(t)}\right) \left(\boldsymbol{y}_i - \hat{\mu}_j^{(t)}\right)'}{n_j^{(t)}}$$

More details for calculating $\hat{\Sigma}_j^{(t)}$ are given by Celeux & Govaert (1995).
For $t = 2, 3....$ repeat the E and M steps in turn until convergence is reached.

Checking Convergence

The EM algorithm's convergence can be examined using different criteria. Lack of progress criteria is a criterion involves checking the difference between the successive log-likelihood values of two iterations as follows:

$$\text{if } \left|\log L(\theta^{(t)}) - \log L(\theta)^{(t-1)})\right| < \text{c} \text{ then convergence is assumed;}$$

where c is a small amount chosen by the user, e.g. 10^{-5}.

The package *mclust* (Scrucca et al., 2016) in R Language is used across this thesis for model-based clustering with Bayesian information criterion (BIC) from *mclust* package to choose the optimal number of clusters (Schwarz, 1978).

2.11.2 Bayesian space-time model for clustering areas based on their disease trends

Napier et al. (2019) proposed a Bayesian space-time model that identifies clusters of similar temporal disease trends. The inference of the model was based on Bayesian statistics using a Metropolis-coupled Markov chain Monte Carlo $(MC)^3$ algorithm. The model is defined as follows:

$$Y_{it} \sim p(y_{it}|\mu_{it}), \quad i = 1,...,n \quad t = 1,...,T,$$
$$g(\mu_{it}) = O_{it} + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + \phi_i + \sum_{s=1}^{S} \omega_{is} f_s(t|\Omega_s).$$

where the study region is divided into n sub-regions with data of T time points, O_{it} is an offset, $\mathbf{x}$ is a vector of covariates with $\boldsymbol{\beta}$ coefficients, and a common spatial effect for all time points ϕ_i modelled using Leroux conditional autoregressive (CAR) prior (Section 2.9). $\sum_{s=1}^{S} \omega_{is} f_s(t|\Omega_s)$ is used for clustering where it is assigned each area to one temporal trend of S. The user specifies the trend functions as constant, linear, known change point, or monotonic cubic spline. This model is suitable when the goal of the study is to identify the groups of areas with similar temporal trends.

2.12 Ward-Like Hierarchical Clustering

Ward-Like hierarchical clustering groups together similar areas based on their characteristics and geographical distance or neighbourhood contiguity by optimising the convex combination $D_\alpha = (1-\alpha)D_0 + \alpha D_1$ (Chavent et al., 2018). The matrix $D_0 = [d_{0,ij}]$ is the dissimilarity matrix of the data features and the matrix $D_1 = [d_{1,ij}]$ is the dissimilarity matrix of geographical distance or neighbourhood contiguity. The adjacency matrix M used where $D_1 = 1_n - M$ with $m_{ii} = 1$, $m_{i,j} = 1$ if the area i and j are neigh-

bour and 0 otherwise. Consider a partition $\mathscr{P}_G^\alpha = (C_1^\alpha, ..., C_G^\alpha)$ in G clusters, where G is chosen from the dendrogram of the dissimilarity matrix D_0. The mixed pseudo inertia I_α of the cluster C_g^α is defined as follows:

$$I_\alpha\left(C_g^\alpha\right) = (1-\alpha)\sum_{i\in C_g^\alpha}\sum_{j\in C_g^\alpha}\frac{w_i w_j}{2\mu_g^\alpha}d_{0,ij}^2 + \alpha\sum_{i\in C_g^\alpha}\sum_{j\in C_g^\alpha}\frac{w_i w_j}{2\mu_g^\alpha}d_{1,ij}^2,$$

where w is the weights of observations (the default is observations are weighted by $1/n$), the weight of the cluster C_g^α is defined as $\mu_g^\alpha = \sum_{i\in C_g^\alpha} w_i$, and $\alpha \in [0,1]$ is the mixing parameter which controls the part of pseudo-inertia due to D_0 and D_1. When the mixing parameter α increases, the geographical homogeneity computed with D_1 increases too, whereas the homogeneity with D_0 decreases (Figure 2.5). One can use appropriate pseudo-within-cluster inertia to find these homogeneities. The proportion of the total mixed pseudo inertia is

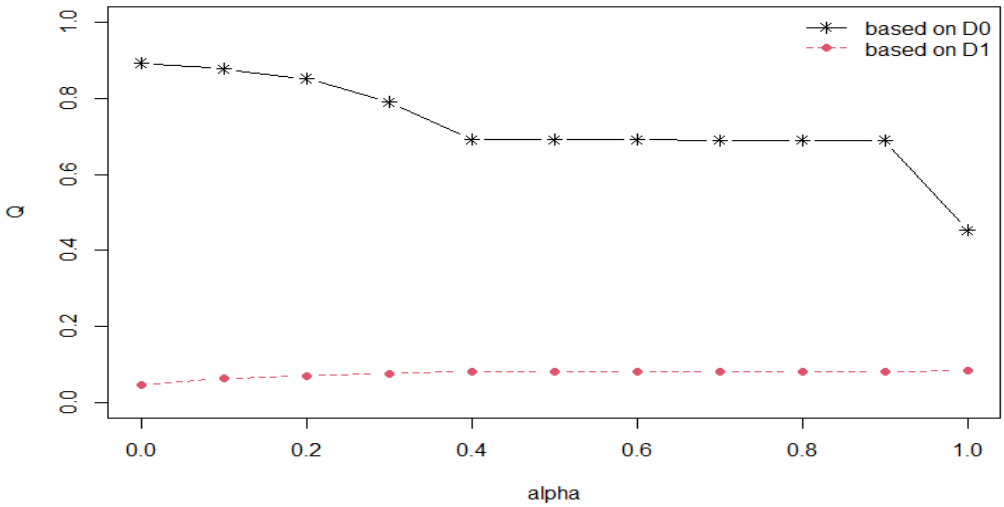
$$Q_\beta\left(\mathscr{P}_G^\alpha\right) = 1 - \frac{W_\beta\left(\mathscr{P}_G^\alpha\right)}{W_\beta\left(\mathscr{P}_1\right)} \in [0,1].$$

where $\beta \in [0,1]$, and the mixed pseudo-within-cluster inertia of $\mathscr{P}_G^\alpha$ is defined as follows:

$$W_\beta(\mathscr{P}_G^\alpha) = \sum_{g=1}^{G} I_\alpha(C_g^\alpha).$$

Then, α can be chosen as a trade-off between the gain of the spatial information and the loss of the data characteristics. As we used the neighbourhood matrix for the dissimilarities in D_1, the value of Q for D_1 will take small values. In this case, one can plot the normalised proportion of explained pseudo-inertias (Figure 2.5).

(i)



(ii)

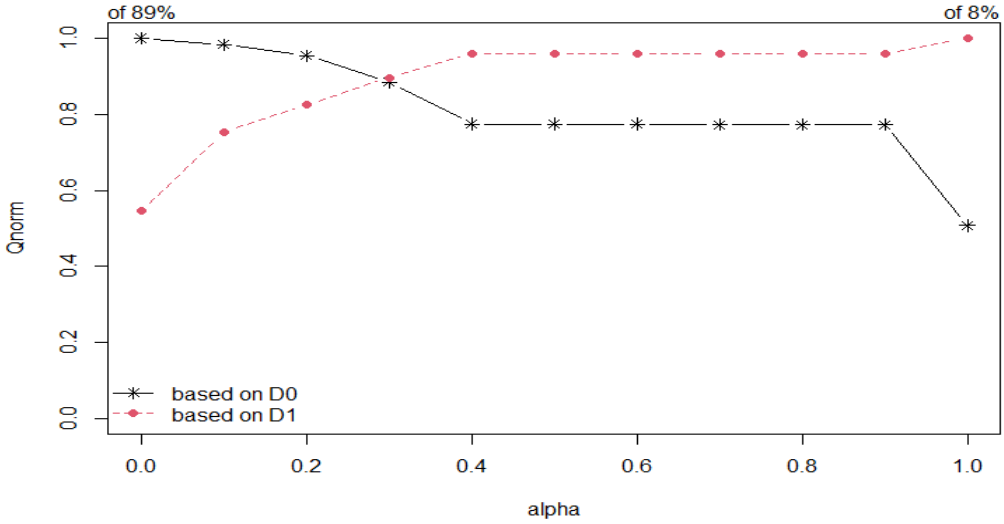


Figure 2.5: Selecting α in Ward-Like hierarchical clustering. (i): proportion of explained pseudo-inertias vs α for D_0 (in a solid black line with *) and D_1 (in dashed red line with •). (ii): normalised proportion of explained pseudo-inertias vs α for D_0 (in a solid black line with *) and D_1 (in dashed red line with •). The normalised plot suggests $\alpha = 0.3$.

2.13 Adjusted rand index for cluster comparison

Adjusted rand index (ARI) proposed by (Hubert & Arabie, 1985) is a common statistical metric used to evaluate a clustering technique's performance. The adjusted Rand index (ARI) is the corrected version of the Rand index, which considers that some agreement between two clusterings can occur by chance. So it, unlike the Rand Index, has an expected value of 0 under random allocation. ARI compares two cluster structures' similarity and takes a value between 0 and 1. A value of 1 indicates perfect agreement between the two cluster structures, whereas 0 indicates that the two cluster structures do not agree. Assume n objects are partitioned into two different cluster structures, $K_C = (K_1, ...K_C)$ and $R_L = (R_1, ..., R_L)$ with c and l clusters respectively. Table 2.1 will explain the Adjusted rand index metric calculation.

Table 2.1: Adjusted Rand Index

	K_1	K_2	...	K_C	sums
R_1	n_{11}	n_{12}	...	n_{1C}	$a_1 = \sum_{i=1}^{C} n_{1i}$
R_2	n_{21}	n_{22}	...	n_{2C}	$a_2 = \sum_{i=1}^{C} n_{2i}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
R_L	n_{L1}	n_{L2}	...	n_{LC}	$a_L = \sum_{i=1}^{C} n_{Li}$
sums	$b_1 = \sum_{j=1}^{L} n_{j1}$	$b_2 = \sum_{j=1}^{L} n_{j2}$	...	$b_C = \sum_{j=1}^{L} n_{jC}$	n

In Table 2.1, n_{ij} represents the number of objects in common between clusters K_C and R_L, and based on the previous table, the adjusted Rand index will be calculated as follows:

$$ARI = \frac{\sum_{j=1}^{L}\sum_{i=1}^{C}\binom{n_{ij}}{2} - \left[\sum_{j=1}^{L}\binom{a_j}{2}\sum_{i=1}^{C}\binom{b_i}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_{j=1}^{L}\binom{a_j}{2} + \sum_{i=1}^{C}\binom{b_i}{2}\right] - \left[\sum_{j=1}^{L}\binom{a_j}{2}\sum_{i=1}^{C}\binom{b_i}{2}\right] \Big/ \binom{n}{2}} \quad (2.13.1)$$

2.14 Label switching

Label switching is a phenomenon that often occurs in mixture models, particularly in Bayesian inference contexts. Various approaches have been proposed to address label switching, including using identifiability constraints in the model specification. Richardson & Green (1997) proposed applying various identifiability constraints, i.e., rearranging the MCMC output differently. They recommended post-processing the MCMC output using different label choices to obtain the component parameters. They proposed selecting labels based on ordering by means, variances, and mixture proportions. The identifiability constraints aims to break the symmetry of the prior distribution, consequently affecting the symmetry of the posterior distribution as well. Ordering the means of the distribution of the components is type of constraint, and it has been used in this thesis. Stephens (2000) developed relabelling algorithms focusing on identifying the permutation of the parameters that minimises a loss function.

2.15 Summary

This chapter provided an overview of the statistical methods used throughout this thesis. It started with a general overview of two major statistical approaches: frequentist and Bayesian statistics. More details of concepts like prior distributions, Bayesian infer-

ence, and techniques such as Markov chain Monte Carlo, including popular algorithms such as the Metropolis-Hastings and Gibbs Sampling were then provided. Moreover, the finite mixture model and a generalized additive model were included here, along with details about splines. Different ways to evaluate and compare multiple statistical models when dealing with the same dataset were discussed. Since this thesis focused on spatial data and clustering methods, there was a special focus on spatial modelling techniques and clustering approaches. Models which will be compared against the proposed models in this thesis were introduced. Lastly, a common statistical measure used in this thesis called the adjusted Rand index was discussed, which is used to evaluate how well clustering techniques perform. The following chapter will outline the background of the coronavirus disease (COVID-19) pandemic, introduce the concept of disease mapping, and provide related literature on disease mapping modelling.

Chapter 3

Disease mapping and Coronavirus disease Data

In the previous chapter, we outlined the statistical theory and methodology, as well as some spatial and spatio-temporal modelling, used throughout this thesis. This chapter will outline the background of the Coronavirus disease (COVID-19) pandemic, which will be studied in this thesis. It also introduces the concept of disease mapping as well as related literature on disease mapping modelling.

3.1 Disease mapping

Disease mapping is one part of the study of statistical epidemiology, where maps are used to depict spatial patterns and illustrate the differences in risk across a region (e.g. a city), which is partitioned into n non-overlapping sub-regions (such as health board areas). The goal is to identify patterns and incidences of high disease risk among the different sub-regions. This disease risk visualisation makes it simpler to compare hazards throughout the region and locate interesting characteristics on the map to help public health officials concentrate on better supporting people in these areas with healthcare

resources.

For the disease mapping data, consider a study region $\mathbf{B}$, which is partitioned into n non-overlapping areal units $\mathbf{B} = \{B_1, ..., B_n\}$. Health data typically consists of disease counts aggregated within these geographical units to maintain patient anonymity. For each areal unit $B_i, i = 1, ..., n,$, the response is collected and denoted by y_i, representing the number, rate, or other summary of disease cases observed within areal unit i. For the n areal units, the responses are $\boldsymbol{y} = (y_1, ..., y_n)$. In this thesis, the study region is the Greater Glasgow and Clyde Health Board, which contains $n = 257$ administrative regions known as Intermediate Zones (IZs), with approximately 1,180,000 people in total and a median population of 4,400 (Health & Data, 2020). This region contains Glasgow, the largest city in Scotland and the surrounding areas (East Dunbartonshire, East Renfrewshire, Glasgow City, Inverclyde, Renfrewshire and West Dunbartonshire). (Figure 3.1).

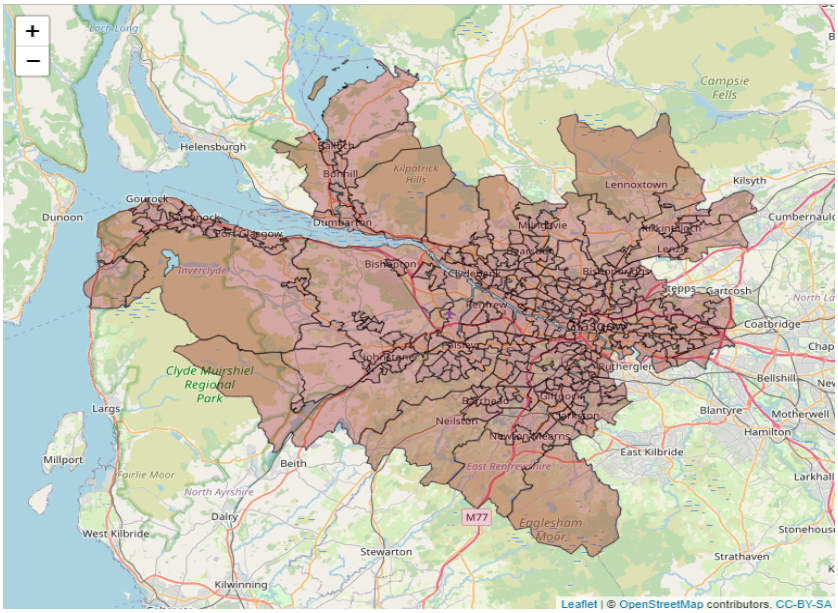


Figure 3.1: Map of the 257 Intermediate Zones (IZs) of the Greater Glasgow and Clyde Health Board.

3.2 Coronavirus disease (COVID-19)

The lives and livelihoods of millions of people throughout the world were in danger due to the global pandemic known as COVID-19 (World Health Organization, 2020*c*). On 31^{st} of December 2019, the World Health Organization (WHO) was informed of cases of pneumonia of unknown aetiology found in Wuhan City in China. The national authorities in China reported 44 cases of novel coronavirus infection in Wuhan City, with some cases in critical condition by the beginning of 2020. Thailand and Japan reported the first and second cases of novel coronavirus (2019-nCoV) outside China, respectively (World Health Organization, 2020*a*). The number of cases in China rose sharply, where the first case was reported in late December, and 11,821 cases were reported after a month. The cases increased to around 82,545 cases by the end of March (World Health Organization, 2020*b*). At the end of January 2020, the WHO Director-General announced that the novel coronavirus outbreak was a public health emergency of international concern (PHEIC), which is WHO's highest level of alarm (World Health Organization, 2020*d*) as many countries started to report new confirmed cases of Coronavirus disease (Figure 3.2) (Mathieu et al., 2020). COVID-19 was identified as an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus with common symptoms of fever, cough, and loss of taste or smell (Menni et al., 2020). As of June 2020, the virus spread rapidly to most countries across the world (Figure 3.2) with around 6,000,000 cases and 400,000 deaths (Mathieu et al., 2020).

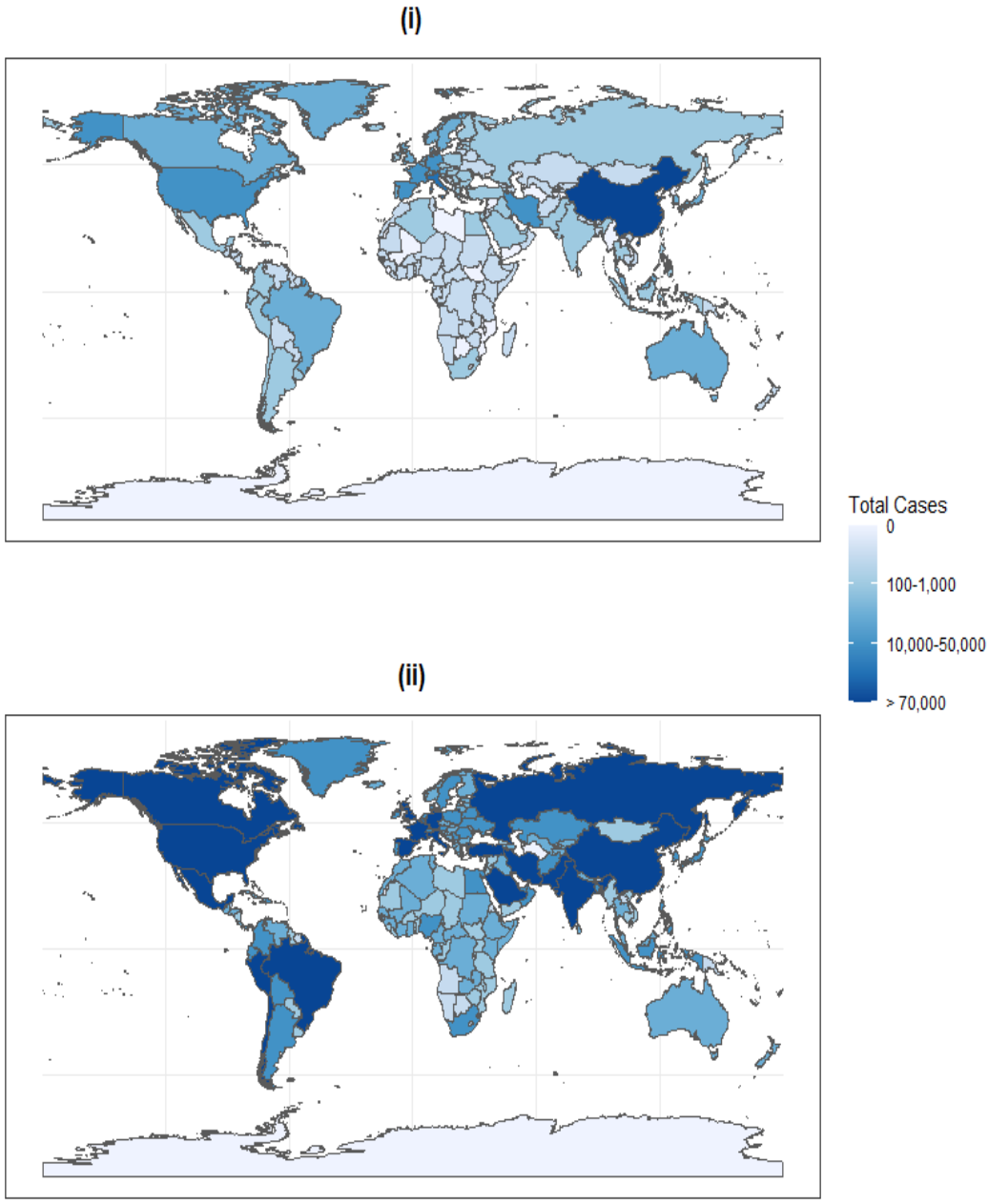


Figure 3.2: World Coronavirus disease cases. Panel (i) shows the Cumulative cases of coronavirus disease for each continent to the end of March 2020. Panel (ii) shows the Cumulative cases of coronavirus disease for each continent by June 2020.

Europe reported a very high number of cases and deaths compared to other countries and on the 13^{th} of March 2020, WHO declared that Europe had become the epicentre of the pandemic (Figure 3.3). In the United Kingdom, the virus was spreading across the country with approximately $5,000,000$ confirmed cases (Figure 3.4) and more than $100,000$ deaths by June 2021.

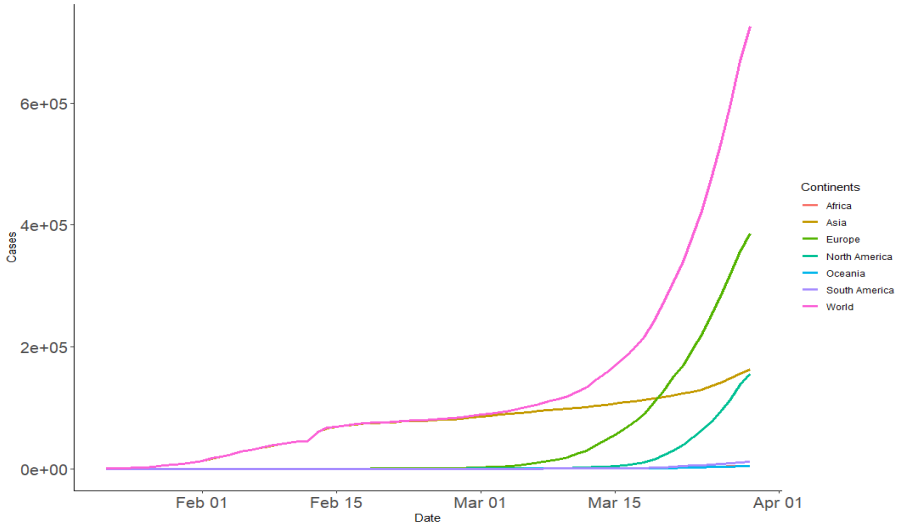


Figure 3.3: Cumulative cases of coronavirus disease for each continent to end of March 2020, with the pink line representing the world cumulative cases of coronavirus disease.

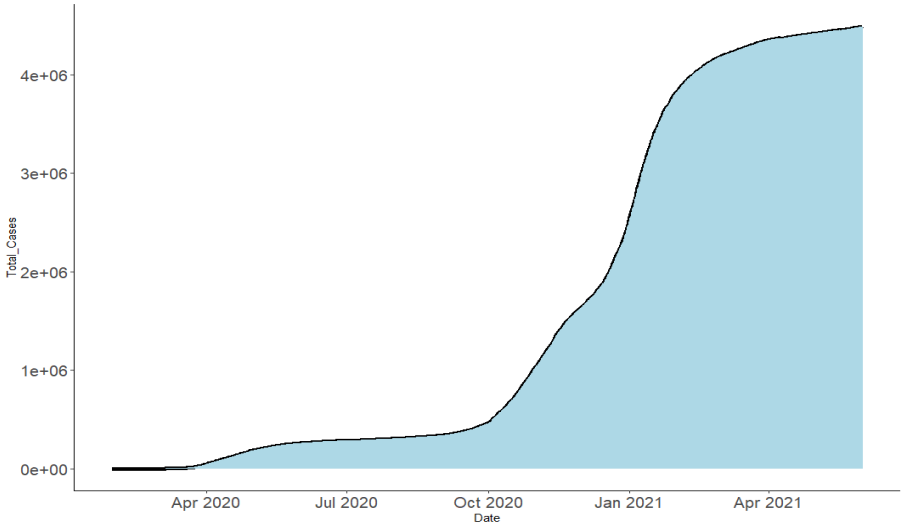


Figure 3.4: Cumulative cases of coronavirus disease in the United Kingdom to June 2021.

On the 1^{st} of March 2020, COVID-19 was confirmed as present in Scotland as the first case was verified (Scottish Government, 2020). According to the NHS, in the Greater Glasgow and Clyde region, at the end of March 2020, there were 679 positive cases, and this total increased to $3,066$ cases with 477 deaths within one month (Health & Data, 2021).

Due to the fast spread of COVID-19 and the high number of cases, a lockdown was necessary. The lockdown refers to actions and restrictions a country has to take to limit the movement of people in case of an emergency situation. Examples of lockdown include stay-at-home orders, closing non-essential businesses such as restaurants and retail stores, schools and universities closing, and travel restrictions. The lockdown aims to keep people safe and secure during crises the country may face, such as the COVID-19 pandemic. The lockdown was nationwide rather than being specific to cities.

Figure 3.5 shows the total number of COVID-19 positive cases over a 7 day period at the beginning of May 2020 for the Greater Glasgow and Clyde Health Board, where cases had increased daily since the first case appeared. During the lockdown, the number of cases decreased to five or fewer, as shown in Figure 3.6. On the other hand, Figure 3.7 shows the 7-day count of cases after the lockdown, where the number of cases increased again after some restrictions were lifted (Health & Data, 2020). During the period of the restrictions, it was a difficult time and caused long-lasting suffering for some people and businesses, but the pressure on the health and social care services was lessened. It is clear how crucial it is to detect high and low-risk areas to help prevent and control the spread of a disease.

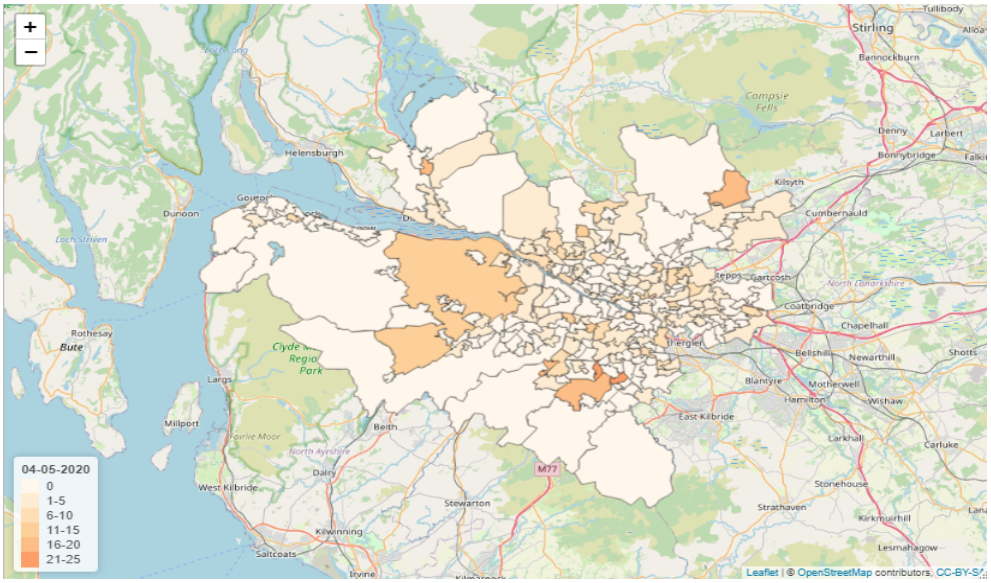


Figure 3.5: The 7-day COVID-19 cases in Greater Glasgow and Clyde in May 2020.

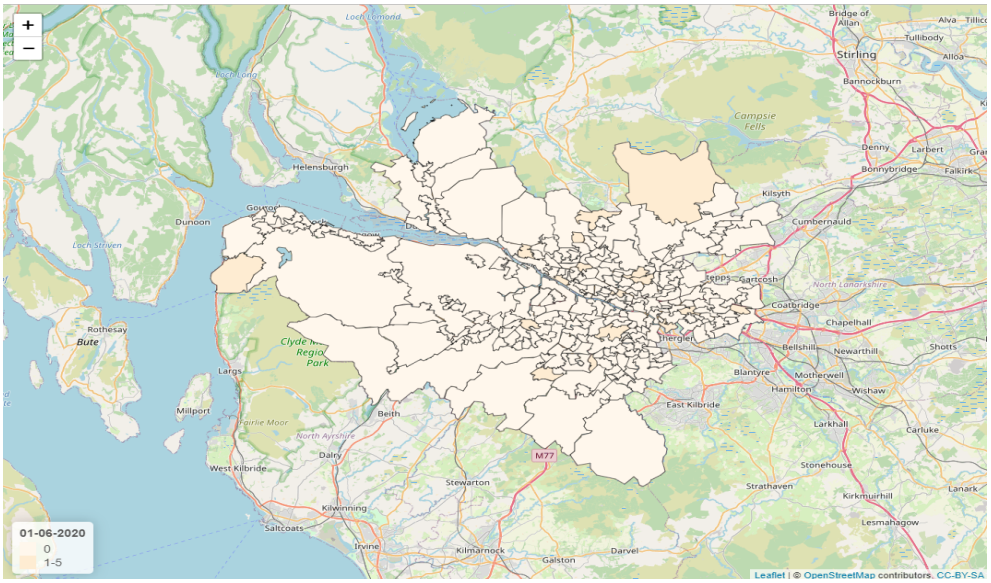


Figure 3.6: The 7-day COVID-19 cases in Greater Glasgow and Clyde in June 2020.

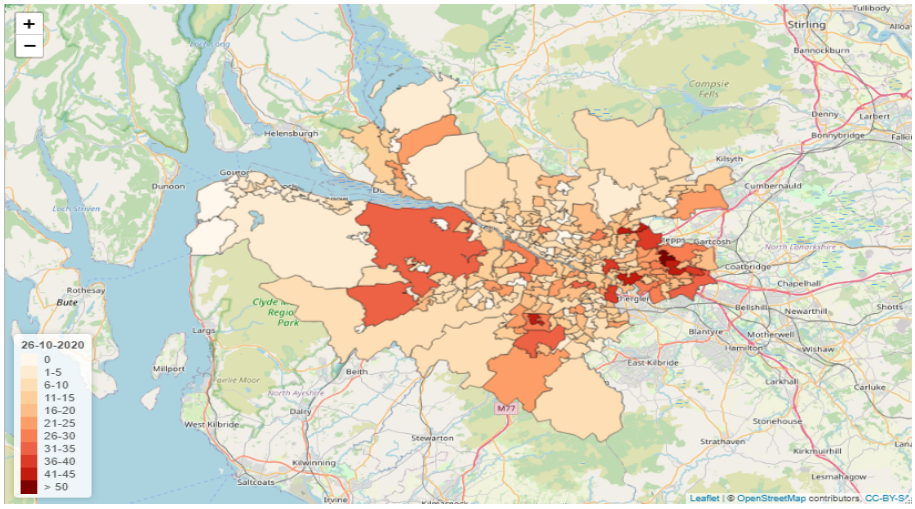


Figure 3.7: The 7-day COVID-19 cases in Greater Glasgow and Clyde in October 2020.

3.3 Spatial clustering

Modelling spatial data is a common challenge in statistical applications as the response variables show spatial dependence where observations spatially close together can be more similar than those further apart (Tobler, 1970). Some statistical approaches have been adopted for dealing with spatial data by identifying clusters of areas that exhibit different levels of disease risk, either high or low risk, compared to their nearby areas. One of the first and most popular methods is scan statistics (Kulldorff, 1997). Although scan statistics are straightforward since they can be implemented with software such as SaTScan, they identify the high-risk clusters but can not estimate the underlying disease risk at the same time. Several Bayesian hierarchical models have been proposed to overcome this drawback.

Knorr-Held & Rasser (2000) developed a non-parametric Bayesian approach based on a reversible jump Markov chain Monte Carlo algorithm (Green, 1995), where all areas are partitioned into a set of contiguous clusters and assume a constant risk within each cluster. Here, a set of areas are chosen to be cluster centres, and after that, the

remaining areal units will be assigned to a cluster based on their closeness to the cluster centres. The model estimates from the data the number of clusters and the location of the cluster centres. Green & Richardson (2002) proposed extended hidden Markov models where the model will assign areas into clusters based on the Potts model with an unknown number of components (Wu, 1982). The inference of this model is also by using the reversible jump Markov chain Monte Carlo algorithm. These approaches are computationally complex and difficult to implement.

Congdon (2007) introduced alternative mixture model schemes for spatial smoothing which considers both continuous and discrete priors. This scheme assumes a priori that an underlying risk for a given area may by similar to, or differ substantially from, neighbouring areas. Charras-Garrido et al. (2012) proposed a method based on a discrete hidden Markov random field (HMRF) for mapping the disease risk clusters. The classes are naturally ordered by their risk levels. Each area will be assigned to one of the risk classes (or clusters), with a penalty that takes into account neighbouring areas in terms of their distance between the risk classes. A Monte Carlo version of the expectation–maximization algorithm was used to estimate the model and post-hoc classification. Considering Bayesian statistics, Wakefield & Kim (2013) proposed a Bayesian version of the Kulldorff (1997) approach. The method defines a list of possible clusters by taking each area individually and continually including the nearby areas in terms of the closest centroid to the centre area until a pre-defined maximum cluster size is achieved. Determining the high and low-risk clusters is done through the data. Note that this method is limited to circle clusters.

Anderson et al. (2014) proposed a two-stage modelling approach. The first stage uses a hierarchical agglomerative clustering algorithm with spatial restrictions to obtain a set of candidate cluster configurations. The second stage is to fit a Poisson log-linear

model for each clustering structure, and the optimal cluster structure is selected using the deviance information criterion (DIC). This approach, therefore, treats identifying the optimal cluster configuration as a model comparison problem. Anderson et al. (2016) proposed a similar approach as Anderson et al. (2014) but estimated disease risk and cluster structure simultaneously in a single model, which is more computationally straightforward. This approach estimates the optimal cluster structure as a parameter within the model rather than via a model comparison.

Using a density-based clustering Santafé et al. (2021) proposed a two-stage method. In the first step, a single cluster structure is estimated using density-based clustering. The next step fits a Bayesian hierarchical spatial model with a single cluster structure. The main aim of this approach is to estimate risks rather than acquire a cluster structure. Wang et al. (2022) proposed a DDW, which stands for data-driven **W**s to adjust spatial dependence in clustering datasets, where **W** is the spatial weight matrix. The method considers the idea of the first low of geographic (Tobler, 1970) and the clustering (or discontinuous). First, the clusters are identified using scan statistics (Kulldorff, 1997), and then DDW is constructed according to the cluster structure. The novel data-driven (DDW) is incorporated into the CAR model.

3.4 Spatio-temporal disease mapping

The field of spatio-temporal modelling for disease risk has become an interesting topic for researchers when disease risk data are available over a period of time. The spatio-temporal modelling identifies changes in the spatial disease risk pattern over time. For analysis of spatio-temporal risk data Bernardinelli et al. (1995) proposed a generalised linear model (GLM) with a linear predictor with a separate linear trend for each area, which is introduced in Section 2.10.1. Waller et al. (1997) introduced an extension of the BYM model (Besag et al., 1991), based on a linear combination of a spatially cor-

related random effect and a non-spatially correlated random effect. Knorr-Held (2000) proposed an approach that included a pair of time-specific effects, unstructured and structured via random effect, and a pair of area-specific effects, unstructured and structured via conditional autoregressive (CAR) model. This model includes an additional term for spatio-temporal interaction.

A generalised additive mixed model (GAMM) has been used in spatio-temporal modelling, an extension of the proposed generalised linear mixed model (GLMM). The model is proposed by MacNab & Dean (2001), combining B-splines smoothing over time effect and the CAR model for the spatial pattern. This model estimates the overall temporal trend by fixed effect B-splines smoothing (de Boor, 1972), whereas the random effect splines are used to separate the small-area trend (more details in Section 2.10.2). Ugarte et al. (2010) proposed a model with P-splines (Eilers & Marx 1996). Separate P-splines were used for the space and time terms and combined these P-splines via tensor product for the space-time effect.

Li et al. (2012) proposed the BaySTDetect model, a model for detecting areas that exhibited unusual trends that differed from the overall region-wide trend. It consists two competing models: the first considers a common temporal trend across the study region and the second model independently estimates time trends for each area. Region-specific probabilities of membership to each competing model are assumed unknown and estimated as part of the MCMC algorithm. Lawson et al. (2012) proposed a model with a spatial effect modelled via a CAR model and a temporal effect modelled via an autoregressive model with the mixture model for space-time interaction, inspired by the idea of mixture model from (Böhning et al., 2000). Rushworth et al. (2014) proposed a model accounting for the space-time autocorrelation using a single set of random effects, where the random effects at time one are modelled via a CAR prior, and the later

time points will be modelled via a CAR prior, but with mean depending on the value of the previous time point's random effects. Napier et al. (2016) proposed a model with a separate independent spatial effect for each time point and an overall temporal term modelled via CAR prior.

Adin et al. (2019) proposed an extension model to the two-stage spatial modelling introduced by Anderson et al. (2014) to two-stage sptaio-temporal modelling. Adin et al. (2019) incorporated temporal trend and space-time interaction. However, this model also faces a computational limitation because it fits multiple Bayesian models separately. Yin et al. (2022) proposes a two-stage modelling approach that takes discontinuities and clusters in the spatio-temporal data into account when estimating the risk. In the first stage, they create a collection of potential neighbourhood matrices which represent the data's range of possible cluster structures. The second stage estimates the optimal structures by considering the neighbourhood matrix as an additional parameter to be estimated within a Bayesian spatio-temporal disease mapping model. The results of the model are less accurate in estimating disease risk and the number of clusters when the expected disease cases in each areal unit is small. Grazian (2023) proposes a novel Dirichlet process incorporating spatial and temporal dependencies in stick-breaking probabilities. The method provides a natural way to test the separability of components and has been shown to provide more accurate predictions when compared with similar existing approaches.

3.5 Summary

This chapter introduced the concept of disease mapping and provides an overview of some studies and research about disease mapping modelling and clustering. Moreover, a brief description of the total cases of COVID-19 globally was given, and maps showing the total number of cases worldwide were offered, giving a bigger picture of the

pandemic's speed. An overview of the COVID-19 cases in the United Kingdom, Scotland, and the Greater Glasgow and Clyde Health Board was then provided which was the region being studied in this thesis. A map of the Greater Glasgow and Clyde Health Board to provide context for subsequent analysis is also provided.

Chapter 4

A Spatially Constrained Poisson Finite Mixture Model

4.1 Introduction

Finding patterns and discrepancies in risk for the disease under research across the study region is one of the critical goals of disease mapping, especially when trying to pinpoint groups of places with a high (or low) risk which are very different from nearby areas. Identifying these sets of high-risk regions will help public health officials identify areas and people who need more health care and enable them to reasonably provide appropriate medical resources. Clustering methodologies provide a potential approach for identifying such regions since they are designed to allocate objects to groups, where the objects in each group share similar features and characteristics and are different from the objects in the other groups. However, when using standard clustering techniques for grouping areal data, such methods will ignore the spatial dependencies between nearby areas.

This chapter aims to introduce a spatial clustering model that incorporates the spatial information of the areal data. We propose a spatially constrained Poisson finite mix-

ture model that estimates the disease risk and identifies high and low-risk clusters for areal data, encouraging spatial contiguity, where areas generally tend to belong to the same cluster as most of their neighbours, except when there is a high difference in some features. The Bayesian approach is used to estimate model parameters, where model inference is carried out using Markov chain Monte Carlo (MCMC) methods using either Gibbs sampling or Metropolis-Hasting steps. The proposed model can identify the high and low risk clusters while taking into account the spatial information and estimating the underlying disease risk in one step. Moreover, it is computationally straightforward, unlike some models which are computationally complex and difficult to implement.

The sections of this chapter are organised as follows. Section 4.2 introduces the proposed model, the spatially constrained Poisson finite mixture model. Section 4.3 outlines the parameter estimation of the model via MCMC. Section 4.4 tests this model against an existing method using simulated data. Section 4.5 outlines the results of the proposed model applied to an application of real data, the Coronavirus disease (COVID-19) count data for the Greater Glasgow and Clyde Health Board. Finally, Section 4.6 summarises the main findings of the proposed approach.

4.2 Spatially Constrained Poisson Finite Mixture Model (SCPFMM)

The basic finite mixture model was introduced in Chapter 2. This model assumes that data are generated from a mixture of g components with unknown parameters vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_g)$ and some mixing proportions $\boldsymbol{\pi} = (\pi_1, ..., \pi_g)$ with $\sum_{j=1}^{g} \pi_j = 1$ and $0 < \pi_j \leq 1$. The model given in equation (4.2.1) groups observations with similar characteristics in the same cluster.

$$f(\mathbf{y_i}|\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{j=1}^{g} \pi_j \phi_j(\mathbf{y_i}|\boldsymbol{\theta}_j). \tag{4.2.1}$$

Areal data has an additional property which must be taken into account, specifically the spatial dependencies between nearby areas. When clustering areal data, areas in the same cluster should have similar characteristics, but the spatial pattern of the region should also be taken into account.

Therefore, a spatially constrained Poisson finite mixture model is proposed here for areal data, which incorporates spatial information into the finite Poisson mixture model's mixing proportions. Sanjay-Gopal & Hebert (1998) proposed a model using the Gibbs prior to incorporate the geographical connections between parameters in image segmentation. Gibbs priors model the spatial relationships between neighbouring areas, which is useful in spatial data where neighbouring areas tend to have similar characteristics. Incorporating Gibbs priors encourages neighbouring areas to be classified within the same group (Geman & Geman, 1984).

The density function of the model is defined as follows:

$$f(\mathbf{y}_i \mid \boldsymbol{\lambda}, \mathbf{p}) = \sum_{j=1}^{g} \left(p_{i,j} \frac{\lambda_j^{\mathbf{y}_i} e^{-\lambda_j}}{\mathbf{y}_i!} \right), \tag{4.2.2}$$

where $p_{i,j}$ denotes the probability of the i^{th} area belonging to the j^{th} group (Sanjay-Gopal & Hebert, 1998). Consider $\mathbf{z}_i$ as a discrete random variable (latent variable) with probability function $p(z_{ij} = 1) = p_{i,j}$. Let $\boldsymbol{z}_i = (z_{i1}, ..., z_{ig})$ be the components membership vector such that $z_{ig} = 1$ if y_i comes from group g and $z_{ig} = 0$ otherwise. Incorporating the latent variables makes sampling easier in terms of the Markov Chain Monte Carlo (MCMC) implementation of the mixture model, where using latent indicator variables usually enables an efficient simulation algorithm that quickly focuses

on the modes of the posterior distribution. The same approach applies to computation using the EM algorithm. The incorporation of these latent (unobserved) variables is done in a way that ensures the likelihood function remains unchanged whilst allowing the identification of each mixture component.

The complete data is $\mathbf{Y}_c = (\mathbf{y}, \mathbf{z})$, and the complete data likelihood function will be defined as follows:

$$L(\lambda, \mathbf{p}, \mathbf{z} \mid \mathbf{y}) = \prod_{i=1}^{N} \left[\prod_{j=1}^{g} \left(p_{i,j} \frac{\lambda_j^{y_i} e^{-\lambda_j}}{y_i!} \right)^{z_{ij}} \right]. \tag{4.2.3}$$

The model will incorporate the spatial information in the individual-specific mixing proportions via a Gibbs density function Markov Random Field based prior of the cluster probability of the i^{th} area, $\mathbf{p}_i = (p_{i1}, ..., p_{ig})$ which is defined as follows:

$$f(\mathbf{p}) = \frac{1}{C} \exp\left(-\beta \sum_{i=1}^{N} V(\mathbf{p}_i) \right), \tag{4.2.4}$$

where C is a normalizing constant, and β is a scale constant which controls the spatial smoothness. The neighbourhood information of the i^{th} area is explained by

$$V(\mathbf{p}_i) = \sum_{m \in N_i} l(u_{i,m}),$$

where N_i represents the set of neighbours around the area i. The function $l(u_{i,m})$ is a penalty function and must be monotonically increasing and also non-negative. Blekas et al. (2005) defined $l(u_{i,m})$ to be as follows:

$$l(u_{i,m}) = \left(1 + u_{i,m}^{-1}\right)^{-1},$$

where

$$u_{i,m} = \|\mathbf{p}_i - \mathbf{p}_m\|^2 = \sum_{j=1}^{g} \left(p_{i,j} - p_{m,j}\right)^2,$$

After incorporating a Gibbs MRF-based prior of $\mathbf{p}_i$ and assuming a Gamma prior with small scale and shape parameter value for the parameters λ_j, where $f(\lambda_j \mid \boldsymbol{a}, \boldsymbol{b}) \sim$ Gamma(a,b), the full posterior distribution will then be

$$\begin{aligned} f(\boldsymbol{\lambda}, \mathbf{p}, \mathbf{z} \mid \mathbf{y}) &\propto L\left(\lambda_j, p_j, z_{ij} \mid \mathbf{y}\right) \times f(\lambda_j \mid \boldsymbol{a}, \boldsymbol{b}) \times f(\mathbf{p}) \\ &\propto \prod_{i=1}^{N} \left(\prod_{j=1}^{g} \left(p_{i,j} \frac{\lambda_j^{y_i} e^{-\lambda_j}}{y_i!} \right)^{z_{ij}} \right) \times \prod_{j=1}^{g} \left(\lambda_j^{a-1} e^{-b\lambda_j} \right) \times \exp\left(-\beta \sum_{i=1}^{N} V(\mathbf{p}_i) \right). \end{aligned} \tag{4.2.5}$$

4.3 Parameter Estimation for the Spatially Constrained Poisson Finite Mixture Model via McMC

The model inference is carried out using the MCMC method. We will use Gibbs sampling for the parameters whose full conditional posteriors follow known distributions, and the Metropolis-Hastings method will be used for the parameters that do not follow a known distribution. The full conditional posterior distributions for each parameter are as follows:

Full conditional posterior distributions for $\boldsymbol{p_i}$

$$f(\boldsymbol{p_i} \mid \mathbf{y}, \mathbf{z}, \boldsymbol{\lambda}) \propto \prod_{j=1}^{g} \left(p_{i,j}\right)^{z_{ij}} \times \exp\left(-\beta \sum_{m \in N_i} \frac{\sum_{j=1}^{g} \left(p_{i,j} - p_{m,j}\right)^2}{\sum_{j=1}^{g} \left(p_{i,j} - p_{m,j}\right)^2 + 1} \right) \tag{4.3.1}$$

A Metropolis Hastings algorithm is used to update the parameter $\boldsymbol{p_i}$, with a proposal function $\boldsymbol{p_i^*}$ drawn from the distribution $\boldsymbol{p_i^*} \sim Dirichlet(\alpha_1, ..., \alpha_g)$, where usually $\alpha_1 =$

$... = \alpha_g = \alpha$. The proposal parameters will be chosen such that we obtain an acceptance rate between 0.2 and 0.6 approximately (Roberts & Rosenthal, 1998). I used an adaptive approach where the proposal distribution variance is reassessed after every 100 MCMC iterations. If the acceptance rate is less than 0.2, the proposal distribution variance was reduced by half; If the acceptance rate is higher than 0.6, the proposal distribution variance was doubled. Otherwise, the variance will not change. For the hyperparameters of the Dirichlet proposal function, I tried symmetric settings, where all hyperparameters were set to the same value, and asymmetric settings, allowing for different hyperparameters such as using the number of areas in each cluster. The symmetric configuration outperformed the asymmetric one, yielding acceptance rates ranging between 0.2 and 0.6.

The Dirichlet distribution has been used as a proposal function for the parameter $\boldsymbol{p}_i$ because it guarantees that the generated values are positive and sum to 1.

Full conditional posterior distributions for λ_j

$$
\begin{aligned}
f(\lambda_j \mid \mathbf{y}, \mathbf{z}, \mathbf{p}, \lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_g) &\propto \prod_{i=1}^{N} \left(\frac{\lambda_j^{y_i} e^{-\lambda_j}}{y_i!} \right)^{z_{ij}} \times \left(\lambda_j^{a-1} e^{-b\lambda_j} \right) \\
&\propto \prod_{i=1}^{N} (\lambda_j^{y_i} e^{-\lambda_j})^{z_{ij}} \times \left(\lambda_j^{a-1} e^{-b\lambda_j} \right) \\
&\propto \lambda_j^{\sum_{i=1}^{N} y_i z_{ij}} \times e^{-\lambda_j \sum_{i=1}^{N} z_{ij}} \times \left(\lambda_j^{a-1} e^{-b\lambda_j} \right) \\
&\propto \left(\lambda_j^{\sum_{i=1}^{N} y_i z_{ij} + a - 1} \right) \left(e^{-\lambda_j (\sum_{i=1}^{N} y_i z_{ij} + b)} \right)
\end{aligned}
$$

$$
f(\lambda_j \mid \mathbf{y}) \sim Gamma \left(\sum_{i=1}^{N} y_i z_{ij} + a, \sum_{i=1}^{N} z_{ij} + b \right) \tag{4.3.2}
$$

Full conditional posterior distributions for z_{ij}

$$(z_{ij})_{j=1}^{g} \sim \text{Multinomial}(1, h_{i1}, ...h_{ig}) \tag{4.3.3}$$

where

$$h_{i,j} = \frac{p_{i,j} f\left(y_i \mid \lambda_j\right)}{f(y_i)}$$

Gibbs sampling will be used to update the parameters λ_j and $\mathbf{z}_i$, where the draws are from Gamma $\left((\sum_{i=1}^{N} y_i z_{ij} + a), (\sum_{i=1}^{N} z_{ij} + b))\right)$ and Multinomial$(1, h_{i1}, ..., h_{ig})$ respectively. To deal with the label switching problem, an order constraint was applied within the MCMC where the values of the means $\boldsymbol{\lambda}$ are in increasing order, i.e. $\lambda_{j-1} \leq \lambda_j \leq \lambda_{j+1}$ (Section 2.14).

The model fitting process involved using the Metropolis-Hastings (MH) algorithm and Gibbs sampling techniques within the R programming language to estimate the parameters of our statistical model. The computational cost was relatively high since the model incorporated Bayesian inference methods involving iterative sampling.

4.3.1 Spatially Constrained Clustering Algorithm

The values of the parameters are sampled using the conditional posterior distribution, where

$$f(\lambda_j \mid \mathbf{y}) \sim Gamma\left((\sum_{i=1}^{N} y_i z_{ij} + a), (\sum_{i=1}^{N} z_{ij} + b))\right) \tag{4.3.4}$$

$$f(\boldsymbol{p_i} \mid \mathbf{y}, \mathbf{z}, \boldsymbol{\lambda}) \propto \prod_{j=1}^{g} \left(p_{i,j}\right)^{z_{ij}} \times \exp\left(-\beta \sum_{m \in N_i} \frac{\sum_{j=1}^{g} \left(p_{i,j} - p_{m,j}\right)^2}{\sum_{j=1}^{g} \left(p_{i,j} - p_{m,j}\right)^2 + 1}\right) \tag{4.3.5}$$

$$(z_{ij})_{j=1}^{g} \sim \text{Multinomial}(1, h_{i1}, ...h_{ig}) \tag{4.3.6}$$

where

$$h_{i,j} = \frac{p_{i,j} f\left(y_i \mid \lambda_j\right)}{f(y_i)}$$

For each iteration, $k = 1, ..., \text{iteration}$, perform the steps outlined below.

Algorithm 1 Spatial constrained Clustering

Initialize $\boldsymbol{\lambda}^0, \mathbf{p}^0, \mathbf{z}^0$
Require for $k = 1, 2, 3, ..., \text{iteration}$ do

Generate a new value λ_j^k from the conditional posterior distribution 4.3.4, $j = 1, ..., g$.

Generate a new value $\mathbf{p}_i^k$ from the conditional posterior distribution 4.3.5, $j = 1, ..., g$.

Generate a new value $\mathbf{z}_i^k$ from the conditional posterior distribution 4.3.6, $j = 1, ..., g$.

The convergence is determined visually by examining the trace plots, which are discussed in more detail in Section 2.2.3.

4.4 Simulation study

This section will illustrate the results of a simulation study using the proposed model outlined in Section 4.2. For determining the number of clusters, we used the deviance information criterion (DIC) (Section 2.8.3) by choosing the number with the lowest DIC value (Spiegelhalter et al., 2002).

4.4.1 Data Generation

We constructed a regular 10×10 grid for our simulation study - these 100 areal units were divided into three spatially contiguous clusters (Figure 4.1). The simulated data

were generated using a Poisson distribution with different means for each cluster. For the initial values of the MCMC, the means $\boldsymbol{\lambda}$, were set to 1 for all groups, while for the proportions $\mathbf{p}$, were set as 1 divided by the number of assumed clusters; $\frac{1}{g}$. The three simulated data scenarios are constructed as follows (Figure 4.2):

- Simulation data set-up 1: $\boldsymbol{\lambda} = (0.5, 20, 50)$-(most separated cluster means).
- Simulation data set-up 2: $\boldsymbol{\lambda} = (0.1, 3, 10)$-(least separated cluster means).
- Simulation data set-up 3: $\boldsymbol{\lambda} = (1, 9, 21)$-(medium separated cluster means).

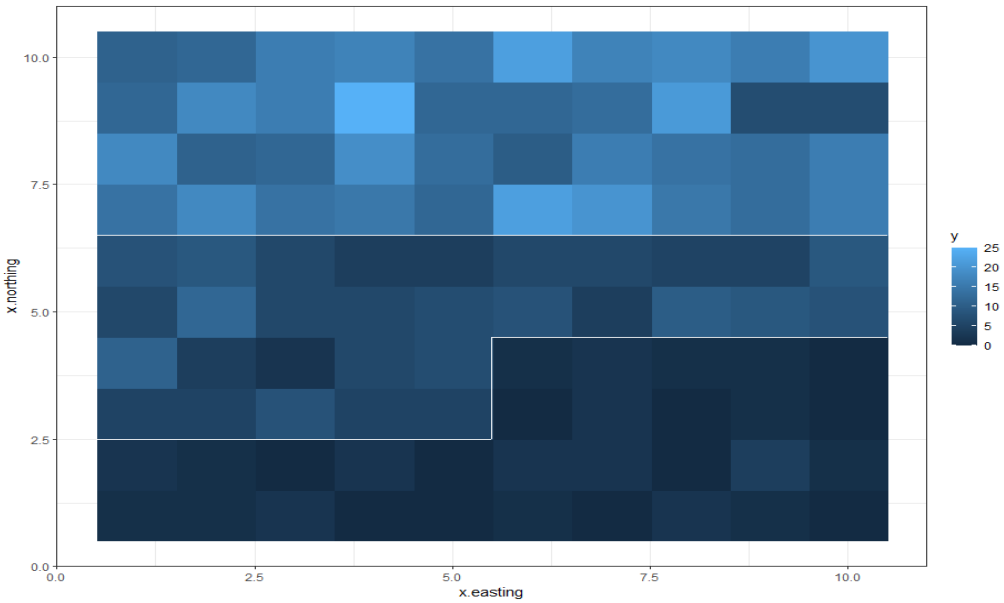


Figure 4.1: Plot of a random simulated data from simulation Set-up 2 with the true three cluster structure boundaries indicated by white lines.

For our simulated data, we chose a gamma distribution as a prior for the parameter λ_j with hyperparameters 0.01 for both shape and scale as a weakly informative prior. For the proposal function of the parameter $\mathbf{p_i}$, a Dirichlet distribution was chosen with initial hyperparameters $\alpha_j = 1 \; \forall j$. The value of the proposal parameters depends on the acceptance rate, explained in Section 4.3. The proposed model was compared to the Ward-Like Hierarchical Clustering (Chavent et al., 2018), which is outlined in Section

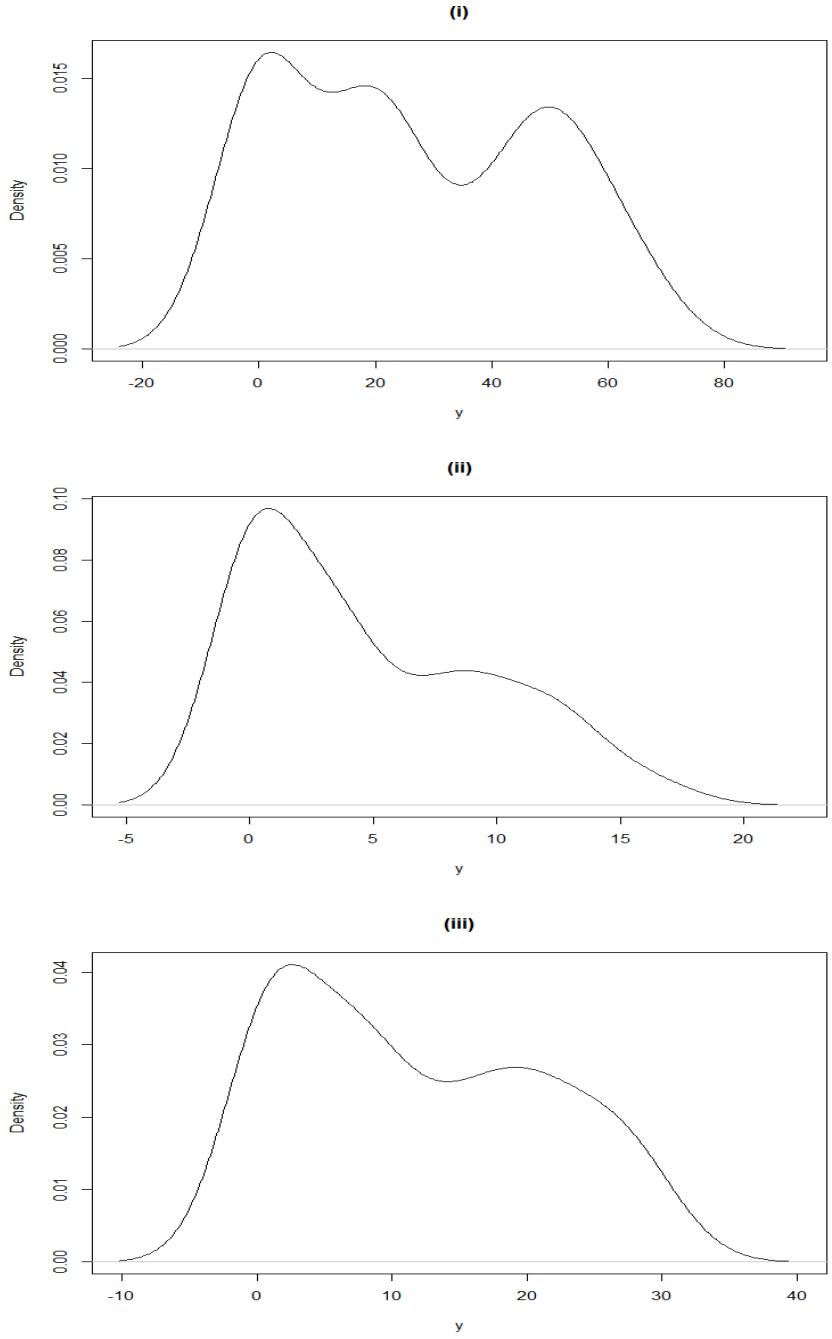


Figure 4.2: Density plots for one set of random simulated data from simulation set-up 1 (i), set-up 2 (ii), and set-up 3 (iii), respectively.

2.12. The proposed model was compared to a Ward-like hierarchical clustering algorithm since it includes spatial constraints and is computationally efficient and straightforward to implement, making it a practical choice. Also, it has equivalencies to one of the most commonly used clustering methods, k-means. Given that the model proposed here incorporated Bayesian inference methods, which can be computationally intensive, 20 replicated datasets was all that was feasible. A smaller subset of five repetitions was implemented initially and gradually increased to 10 and then 20, and we observed consistent average results across the repetitions. Table 4.1 illustrates the results of twenty simulated data sets.

4.4.2 Results

Table 4.1 illustrates the median number of clusters identified by each model. The proposed SCFPMM model selects the number of clusters based on the minimum DIC, where the g range was from 2 to 5 clusters. Also, it shows the median adjusted Rand Index (ARI) (Rand, 1971) comparing the model cluster structure to the true classification for the three simulated data set-ups, where the ARI will get close to 1 when the model is very similar to the true classification (Section 2.13). Table 4.2 illustrates the number of clusters for all simulated datasets across the twenty replicated datasets. The number reported in each table cell is the number of simulations that estimated a specific number of clusters 2, 3 or 4. Figure 4.3 shows boxplots of the adjusted Rand Index for all simulated data sets. For simulation set-up 1, the median number of clusters estimated for both models is 3 with a high adjusted Rand Index, which indicates a good match between the clustering and the true classification. Although both models estimated a correct number of clusters on average, our proposed model has no variability in the ARI values, as seen in the top panel of Figure 4.3. In simulation set-up 2, our model performs better than ward-like hierarchical clustering with a high median ARI

equal to 0.91 compared to a median ARI equal to 0.78 for the Ward-like hierarchical clustering method. Even though both models result in a median number of clusters equal to 3, the middle column of Table 4.2 indicates that the proposed model better estimates the correct number of clusters. Specifically, it accurately estimates the number of clusters in 19 out of 20 replicated simulated datasets. In contrast, the Ward-like hierarchical clustering method achieves this in only 12 out of 20 replicated datasets, with the proposed model exhibiting low variability in the Adjusted Rand Index (ARI) values (Figure 4.3 (ii)). The Ward-like hierarchical clustering in simulation set-up 3 has a median ARI value equal to 0.89, which is less than the median ARI value of our proposed model, which is equal to 0.97 with low ARI value variability (Figure 4.3 (iii)). Once again, our proposed model accurately identifies the number of clusters for all the replicated datasets, whereas the Ward-like hierarchical clustering method only achieves this in 14 out of 20 cases (Table 4.2 third column). Comparing the two methods, the SCPFMM could detect the correct number of clusters and achieve a higher ARI than the other method.

Table 4.1: Median Results of Comparing Cluster Models Fit to Different Simulation Data Set-ups specified in Section 4.4.1.

	Simulation Set-up 1	Simulation Set-up 2	Simulation Set-up 3
SCPFMM [a] No.Cluster	3	3	3
ClustGeo [b] No.Cluster	3	3	3
SCPFMM [a] ARI	1	0.91	0.97
ClustGeo [b] ARI	1	0.78	0.89

[a] Spatially Constrained Poisson Finite Mixture Model
[b] R package for Ward-like hierarchical clustering

Table 4.2: Summary of the estimated number of clusters under each simulation set-up. The number reported in each table cell is the number of simulations that estimated a specific number of clusters (2, 3 or 4).

	Simulation Set-up 1	Simulation Set-up 2		Simulation Set-up 3	
No. clusters / Model	**3**	**2**	**3**	**4**	**3**	**4**
SCPFMM [a]	20	1	19	0	20	0
ClustGeo [b]	20	0	12	8	14	6

[a] Spatially Constrained Poisson Finite Mixture Model
[b] R package for Ward-like hierarchical clustering

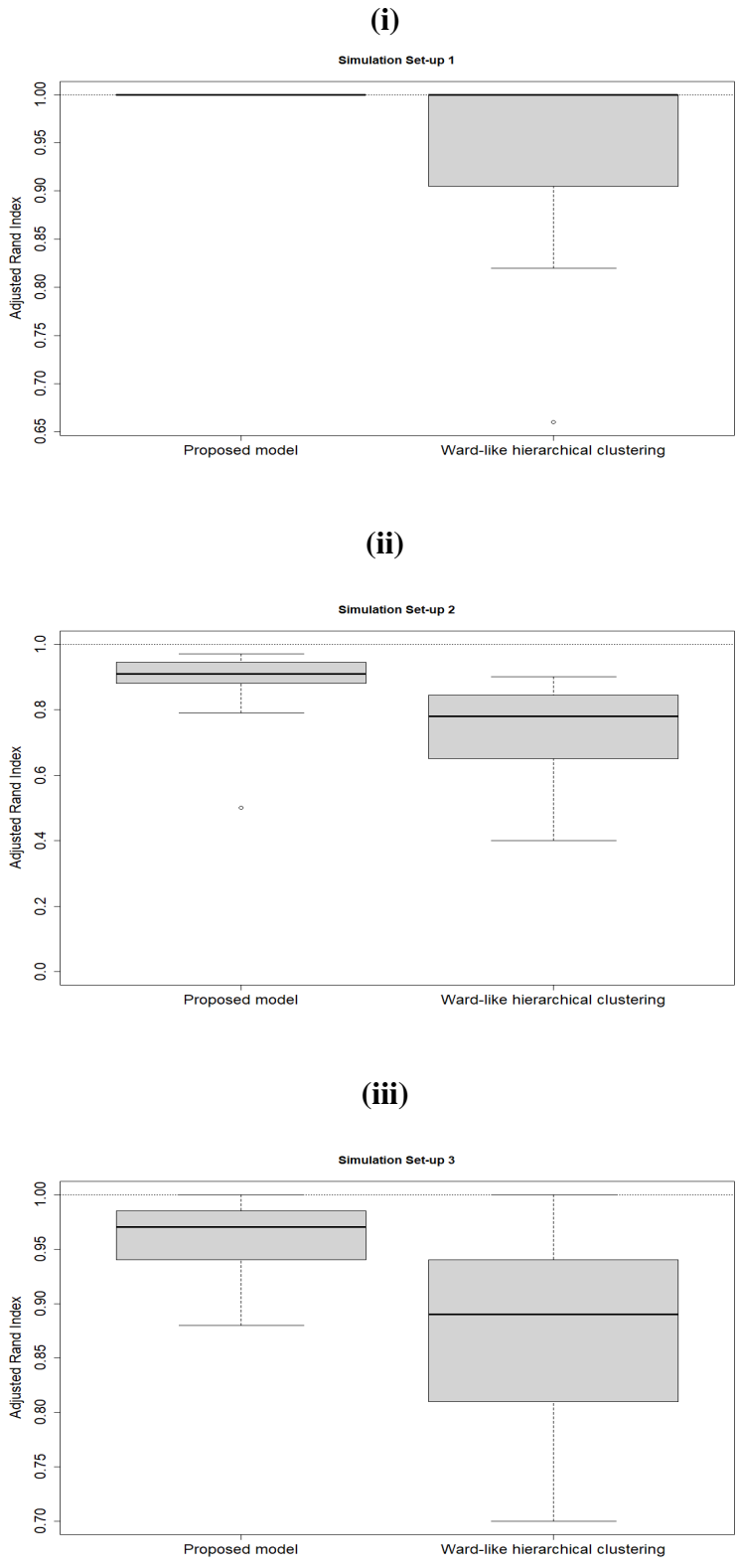


Figure 4.3: Summary of the Adjusted Rand Index obtained under each simulation set-up. The top panel shows a boxplot for simulation set-up 1. The middle panel shows a boxplot for simulation set-up 2, and the bottom panel shows a boxplot for simulation set-up 3. The dotted lines represent a perfect match between the clustering and the truth with a value of 1.

4.5 Application of SCPFMM to COVID-19 data

This section applies our spatially constrained Poisson finite mixture model to Coronavirus disease (COVID-19) cases in Glasgow and its surrounding regions (Health & Data, 2020). The study area is the Greater Glasgow and Clyde Health Board area, which includes the river Clyde estuary in the west and the city of Glasgow in the east, the largest city in Scotland. The health board of this region is split into $n = 257$ administrative regions known as Intermediate Zones (IZs) as displayed in Figure 3.1. The response data, $\mathbf{y} = (y_1, ..., y_n)$, are the 7-day count of COVID-19 cases in the last week of October 2020 for each of the 257 IZs, where y_i represents the 7-day count of COVID-19 cases in area i (Figure 4.4). This particular time point was selected because it coincided with the availability of COVID-19 testing and heightened public awareness of the associated symptoms.

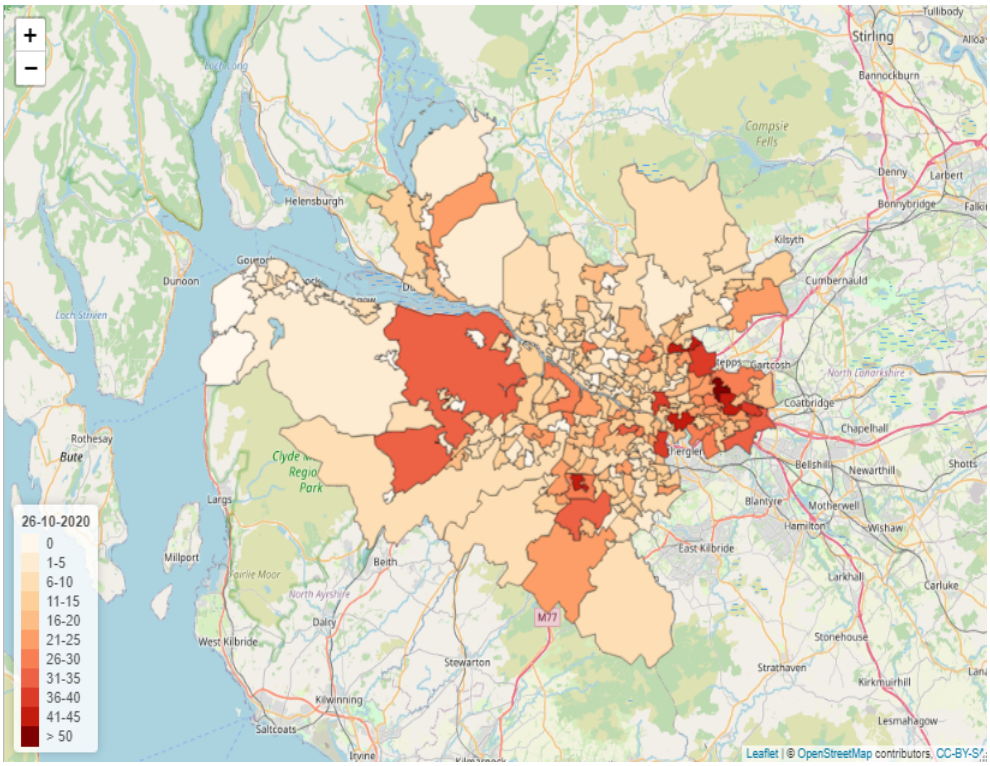


Figure 4.4: The 7-day cases in Greater Glasgow and Clyde (26-10-2020).

4.5.1 Results

We fit the proposed model described in Section 4.2 to the COVID-19 data with different numbers of clusters between 2 to 20, where Figure 4.5 shows the DIC values for these models. The model with four clusters has the lowest DIC value, which we will select as number of clusters for the data. The clusters are plotted on the map where cluster number 1 for the low-risk areas and cluster number 4 for the high-risk areas, as shown in Figure 4.6. The posterior summary for the four clusters is as follows. The first cluster has a mean of around 0 with a standard deviation of 0.004. The 95% credible interval ranges from 0 to 0.002. This cluster, which has a mean of 0 and a narrow range of variability, could represent areas with minimal to no COVID-19 cases. The second cluster has a moderate mean of 6.80 and a standard deviation of 0.36 with a 95% credible interval from 6.1 to 7.5, suggesting a relatively narrow spread of values around the mean. Cluster 2 represents areas with a modest level of COVID-19 transmission. Cluster 3 exhibits higher mean values around 13.90 with a standard deviation of 0.62, and the 95% credible interval from 12.73 to 15.15 suggests a wider spread of values compared to Cluster 1 and 2. The regions in Cluster 3 may have experienced outbreaks leading to a higher infection prevalence than Clusters 1 and 2. Finally, with a mean of 28.07, a standard deviation of 1.11, and a 95% credible interval from 26.13 to 30.20, cluster 4 represents areas with the highest COVID-19 cases. These areas may be experiencing an outbreak or community spread, necessitating urgent interventions and resource allocation. The high-risk cluster included Renfrewshire, East Renfrewshire, and Easterhouse. The model effectively accounts for spatial dependencies, as evidenced by the identification of distinct clusters. It works well in clustering neighbouring areas with similar characteristics but does not produce fully contiguous clusters. For example, Renfrewshire and Easterhouse are in cluster 4, yet they do not share a border. We could consider incorporating additional contiguity penalty terms into the modelling approach if a study objective is to enforce rather than encourage contiguity among the identified

clusters.

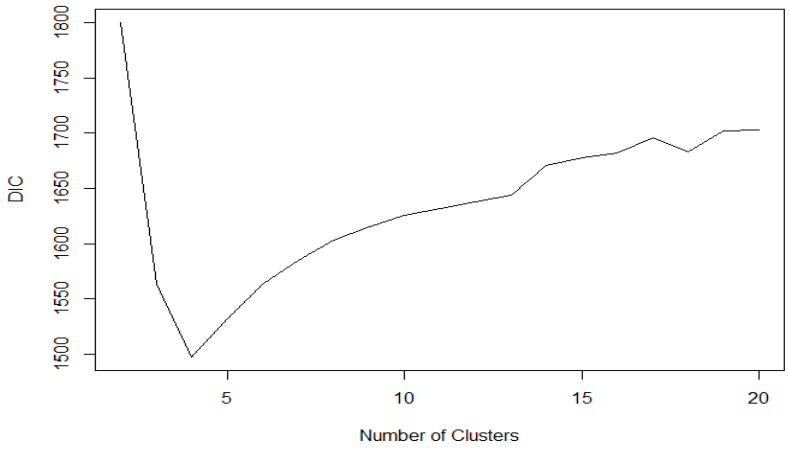


Figure 4.5: Plot of the Deviance Information Criterion for models with between 2 to 20 clusters.

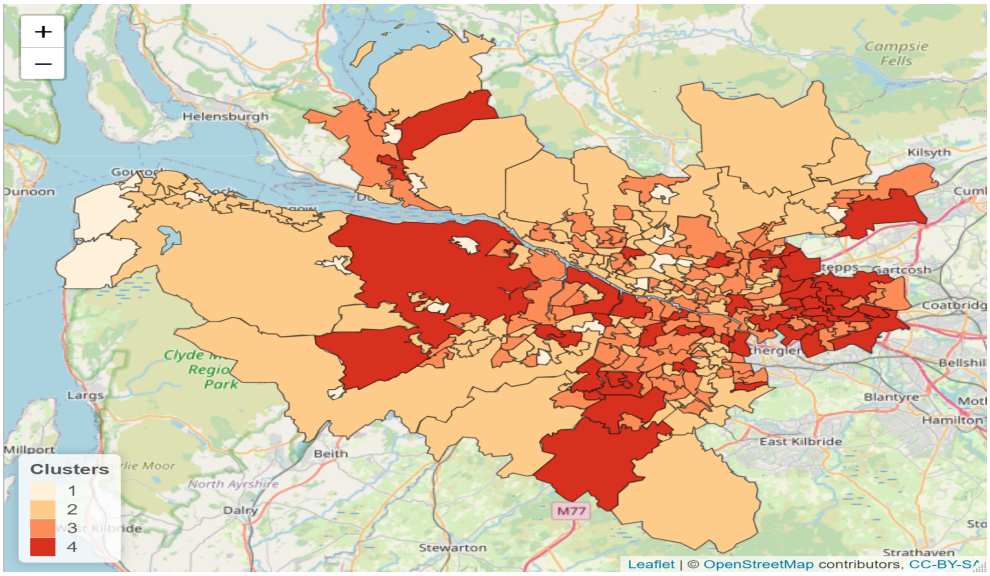


Figure 4.6: A map of Greater Glasgow and Clyde with the SCPFMM estimated clusters.

4.6 Summary

In this Chapter, we have proposed a new approach for clustering spatial count data using Bayesian statistics. The model considers the spatial information of the areal data, where areas belong to the same cluster as most of its neighbours. The clusters estimated by the proposed model take into account both geographic features and observed disease counts. The model incorporated the spatial information through the use of a Gibbs prior combined with a Poisson finite mixture model. In Section 4.4.2, we presented the simulation study results that show our model performs better than the Ward-Like Hierarchical Clustering model (Chavent et al., 2018), in terms of the Adjusted Rand Index, where the proposed model had a high Adjusted Rand index value for the three different simulated data scenarios. In Section 4.5, the proposed model was applied to the COVID-19 data for the Greater Glasgow and Clyde Health Board. The model identifies four clusters for the study areas with high-risk areas, including Renfrewshire, East Renfrewshire, and Easterhouse. As the clusters produced are not fully contiguous, adding additional contiguity penalty terms could be considered.

Chapter 5

Spatio-temporal modelling and clusters detection

5.1 Introduction

The model presented in Chapter 4 aims to detect groups of areal units with similar disease risk at a specific time point. However, disease risk data are available at repeated time points over long periods, and it is crucial to understand how disease risk changes over these periods and particularly identify the areas with increasing disease risk over time to help health authorities. This could allow health authorities to concentrate on better supporting people in high-risk areas with healthcare resources. The field of spatio-temporal modelling focuses on estimating disease risk over a period of time. Clustering the spatio-temporal data has also become more popular. Section 3.4, gave some literature review about approaches to spatio-temporal disease mapping.

This chapter proposes an approach for estimating and clustering spatio-temporal disease risk data. Our goal is to identify clusters of areas that share a similar trend over a period of time. Clustering the interaction coefficients of the p-splines of the interaction between space and time in epidemiological spatial-temporal data represents a new

contribution to clustering the spatial-temporal data field. Overall, the approach appears promising for analyzing spatial-temporal data, offering flexibility, interpretability, and the potential to uncover spatial-temporal patterns.

The chapter is organised as follows. Section 5.2 presents the proposed approach of estimating and clustering spatio-temporal data along with details of the model estimation. Section 5.3 illustrates the efficacy of the new approach via a simulation study and compares it to a competing spatio-temporal model. In Section 5.4, the proposed approach is applied to the Coronavirus disease (COVID-19) count data for the Greater Glasgow and Clyde Health Board during the time period from August 2020 to October 2021, which was introduced in Section 3.2. Finally, Section 5.5 summarises the main findings and the advantages of the proposed approach.

5.2 Methodology

This section proposes a two-step approach for estimating spatio-temporal disease risk and identifying the clusters of the spatio-temporal data. In the first step, a generalised additive mixed model is fitted to the spatio-temporal data inspired by the MacNab & Dean (2001) model. P-splines are used for the smoothing terms, which contain B-splines and discrete differences penalised on their coefficients to control "wiggliness", that is, the model balances the trade-off between the smoothness and overfitting of the data (more details are given in Section 2.7). The second step is to apply a model-based clustering model (Fraley & Raftery, 2002) on the coefficients of the smoothing terms of the space-time interaction to estimate a cluster structure model.

5.2.1 Proposed spatio-temporal model

We propose a spatio-temporal generalised additive mixed model for count data outcomes that includes spatially correlated random effects via the conditional autoregressive (CAR) model (Section 2.9.2), fixed effects P-splines smoothing for modelling the overall temporal trends, and random effects P-splines for modelling the space-time interaction (Section 2.7.2), which provide flexibility in capturing non-linear relationships in the spatio-temporal data.

Let y_{it} be the observed number of cases for the i^{th} area at the time t with $i = 1,2,...,n$ and $t = 1,2,...,T$. Note that we assume areal unit areas stay the same over time and have no missing data.

The general formula will be as follows:

$$\begin{aligned} y_{it} &\sim \text{Poisson}(\mu_{it}) \quad i = 1,...,n, \quad t = 1,...,T, \\ \log \mu_{it} &= \log n_{it} + \log \varpi + A(t) + \phi_i + \beta_i(t) \end{aligned} \tag{5.2.1}$$

where n_{it} represents the population in the i^{th} area at the time t, ϖ is the mean rate over the time period, and all areas, and ϕ_i is the spatial random effect. The fixed effects smoothing function of t is expressed via $A(t)$, and the random effects smoothing function of the space–time interaction is expressed via $\beta_i(t)$.

After applying a P-spline for the fixed and the random temporal effects, the model will be defined as follows:

$$\log \mu_{it} = \log n_{it} + \log \varpi + S_0(t) + \phi_i + S_i(t) \tag{5.2.2}$$

where ϕ_i is modelled using an intrinsic CAR model, $\boldsymbol{\phi} = (\phi_1, ..., \phi_i) \sim N(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W})^-)$, where $\mathbf{Q}(\mathbf{W})$ is a precision matrix, $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$, and $\mathbf{W1}$ is a vector containing the number of neighbours for each areal units. $\mathbf{W}$ is a neighbourhood matrix, which is specified as $w_{ij} = 1$ if the two areas are sharing a common border and $w_{ij} = 0$ if they do not share a common border, and τ is a scalar precision parameter. $S_0(t) = \sum_{k=1}^{K} p_k(t)\beta_{0,k}$ is a set of P-splines for modelling the overall potentially non-linear temporal trend and $S_i(t) = \sum_{k=1}^{K} p_k(t)B_{i,k}$ is a P-spline smoothing for the time of area i which allow for nonlinear area-specific deviations. The number of the P-spline basis functions is denoted by K and p_k, $k = 1, ..., K$, represents the basis function. The fixed effect vector is $\mathbf{a} = (\log \varpi, \beta_{0,1}, ..., \beta_{0,K})$, whereas the random effect is $\mathbf{b} = (\phi_1, ..., \phi_i, ..., \phi_n, B_{1,1}, ..., B_{1,K}, ..., B_{i,k}, ..., B_{n,K})$. The proposed model was fitted using the `mgcv` package (Wood, 2017). After fitting this model to the spatio-temporal data, we will partition the areas by applying the model-based clustering algorithm (Fraley & Raftery, 2002) on the estimated P-splines coefficients for the random effect of the space–time components $S_i(t)$.

5.2.2 P-splines fitting

The model coefficients are estimated by penalized maximum likelihood estimation, where the penalized iteratively re-weighted least squares(PIRLS) algorithm is used to maximize the penalized log-likelihood. Let $S_0(t)$ be defined as follows:

$$S_0(t) = \sum_{k=1}^{K} p_k(t)\beta_{0,k},$$

where $\beta_{0,k}$ are coefficients to be estimated. The penalty for the smooth term is written as $\boldsymbol{\beta}^\top \mathbf{P}_0 \boldsymbol{\beta}$, where $\mathbf{P}_0$ is a penalty matrix. The penalized log-likelihood is defined as follows:

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \lambda_0 \mathbf{P}_o \boldsymbol{\beta}/2,$$

The PIRLS algorithm steps are as follows:

1. Initialize $\hat{\mu}_i = y_i + \delta_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$, where δ_i is a small constant value.

   Iterate the next two steps until convergence is achieved.

2. Compute $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$, and $m_i = 1/(g'(\hat{\mu}_i)^2 V(\hat{\mu}_i))$.

3. Find $\hat{\boldsymbol{\beta}}$ to minimize the weighted expression

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_M^2 + \boldsymbol{\beta}^\top \lambda_o \mathbf{P}_o \boldsymbol{\beta},$$

where $V(\mu_i) = var(y_i)$, $\mathbf{M}$ is a diagonal matrix with diagonal elements m_i. The restricted maximum likelihood (REML) and the performance-oriented iteration are used to estimate and select the smooth parameter. The REML criterion with QR decomposition on $\mathbf{X}$, $\sqrt{\mathbf{M}}\mathbf{X} = QR$, to improve the REML has the flowing form:

$$\frac{\|Y - R\hat{\beta}_{\lambda_0}\|^2 + \hat{\beta}_{\lambda_0}^\top \lambda_0 \mathbf{P}_0 \hat{\beta}_{\lambda_0} + r}{2\Phi} + \frac{\log|R^\top R/\Phi + \lambda_0 \mathbf{P}_0/\Phi| - \log|\lambda_0 \mathbf{P}_0/\Phi|_+}{2}$$

where where Q has orthogonal colums and R is upper triangular, $Y = Q^\top \sqrt{\mathbf{M}}\mathbf{z}$, $r = \|\sqrt{\mathbf{M}}\mathbf{z}\|^2 - \|Y\|^2$, and $\hat{\beta}_{\lambda_0} = \text{argmin}_\beta \|Y - \mathbf{R}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^\top \lambda_0 \mathbf{P}_0 \boldsymbol{\beta}$. The Performance-oriented iteration is to select the smooth parameters on the working model at each step of the PIRLS.

For $S_i(t) = \sum_{k=1}^{K} p_k(t) B_{i,k}$, the penalty term is $\mathbf{B}^\top \lambda \mathbf{P}\mathbf{B}$ with $\mathbf{B}$ is a vector of basis coefficients and $\mathbf{P}$ is a penalty matrix. As the true function is believed to be fairly smooth, we express our belief using a Bayesian manner with a prior, $B \sim MVN(\mathbf{0}, (\lambda \mathbf{P})^{-1})$. we estimate the smooth function $S_i(t)$ by minimizing

$$\|\mathbf{y} - S_i(t)\|^2 + \mathbf{B}^\top \lambda \mathbf{P}\mathbf{B}$$

The proposed model uses the `mgcv` package with the $s()$ function. This $s()$ function defines the smooth term within the generalized additive mixed model, and we use the

argument (bs="ps") for the P-spline basis function. Wood (2017) provides more details on the generalized additive models and their inference. In terms of choosing the optimal number of basis dimensions, we check the RMSE.

In the second step, the model-based clustering approach, which is introduced in Section 2.11.1, is applied on the estimated P-splines coefficients for the random effect of the space–time components $S_i(t)$ using the `mclust` package (Scrucca et al., 2016). Within the `mclust` framework, the optimal number of clusters is selected using the negative Bayesian information criterion (Section 2.8.2), where the number of clusters with the highest BIC value is preferable.

5.2.3 Model-based clustering of splines coefficients

Let $\mathbf{B}_i = (B_{i,1}, ..., B_{i,K})$ represent the estimated P-splines coefficients for area i. The finite mixture model with g groups is defined by

$$f(\mathbf{B}_i) = \sum_{j=1}^{g} \pi_j f_j(\mathbf{B}_i|\mu_j, \Sigma_j) \tag{5.2.3}$$

where π_j is is the prior probability of membership of group $j, j = 1, ..., g$, and $f_j(\mathbf{B}_i|\mu_j, \Sigma_j)$ is the density of a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j.

The likelihood function is:

$$L(\mu, \Sigma, \pi|\mathbf{B}) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j f_j(\mathbf{B}_i|\mu_j, \Sigma_j) \tag{5.2.4}$$

The expectation maximization (EM) algorithm 2.11.1 is used for estimating mixture model parameters. We consider the complete dataset to be viewed as $(\mathbf{B}_i, \mathbf{z}_i)$ where z_{ij}

is a binary indicator variable that equals 1 if the observation i belongs to cluster j, and 0 otherwise.

The likelihood of the complete data will be defined as

$$L_{\text{complete}}(\mu, \Sigma, \pi, \mathbf{z}|\mathbf{B}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\pi_j f_j(\mathbf{B}_i|\mu_j, \Sigma_j)\right)^{z_{ij}} \tag{5.2.5}$$

where:

- π_j is the prior probability of membership of group j, where $j = 1, ..., g$.
- $f_j(\mathbf{B}_i|\mu_j, \Sigma_j)$ is the density of a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j.
- z_{ij} is a binary indicator variable that equals 1 if the observation i is belong to cluster j, and 0 otherwise.

For $v = 2, 3....$ repeat the E and M steps in turn until convergence is reached.

E-Step (Expectation)

$$\hat{z}_{ij}^{(v)} = \frac{\hat{\pi}_j^{(v-1)} f_j(\mathbf{B}_i|\hat{\boldsymbol{\theta}}_j^{(v-1)})}{\sum_{l=1}^{g} \hat{\pi}_l^{(v-1)} f_l(\mathbf{B}_i|\hat{\boldsymbol{\theta}}_l^{(v-1)})}$$

M-Step (Maximization)

$$n_j^{(v)} = \sum_{i=1}^{n} \hat{z}_{ij}^{(v)}$$

$$\hat{\pi}_j^{(v)} = \frac{n_j^{(v)}}{n}$$

$$\hat{\mu}_j^{(v)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(v)} \mathbf{B}_i}{n_j^{(v)}}$$

$$\hat{\Sigma}_j^{(v)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(v)} \left(\mathbf{B}_i - \hat{\mu}_j^{(v)}\right) \left(\mathbf{B}_i - \hat{\mu}_j^{(v)}\right)^{\top}}{n_j^{(v)}}$$

More details for calculating $\hat{\Sigma}_j^{(v)}$ under various restrictions are given by Celeux & Govaert (1995) and Scrucca et al. (2016). The optimal number of clusters is selected using the negative Bayesian information criterion (Section 2.8.2), where the number of clusters with the highest BIC value is preferable.

5.3 Simulation study

5.3.1 Aim

A simulation study is carried out in this section to determine the effectiveness of the proposed model outlined in Section 5.2 to identify groups of areas based on sharing temporal trends. The proposed model is compared to the Bayesian space–time model for clustering areas based on their disease trends (STCARclustrends) outlined in Section 2.11.2, presented by Napier et al. (2019), which group areas together that share similar temporal trends. It is straightforward to implement in the R language using the function the ST.CARclustrends() function within the CARBayesST package version 4.0 (Lee et al., 2018). Below is a summary of the simulation study's findings.

5.3.2 Data Generation

The simulated data were generated for $n = 121$ areas and $T = 100$ time points using a generalized additive mixed model with Poisson distribution with three clusters where

adjacent cells tend to have similar distributions (Figure 5.1).

The data were generated using the following model:

$$\begin{aligned} y_{it} &\sim \text{Poisson}(\mu_{it}) \quad i = 1,...,121, \quad t = 1,...,100, \\ \log \mu_{it} &= S_0(t)^c + \phi_i + S_i(t) \quad c = 1,2,3, \\ \phi_i &\sim N(0, Q^{-1}). \end{aligned} \tag{5.3.1}$$

We generated c: three different temporal trends using cubic P-splines $S_0(t)^c$ outlined in detail in Section 2.7.2. We added random effects generated from a multivariate Gaussian distribution with a spatially correlated precision matrix, $Q = \text{diag}(W1) - W$, where W is a neighbourhood matrix, which is specified as $w_{ij} = 1$ if the two areas are sharing a common border and $w_{ij} = 0$ if they do not share a common border. The intrinsic CAR (ICAR) model, which is defined in Section 2.9.2, was used for the spatial effects (Besag et al., 1991). $S_i(t)$ is a P-spline smoothing for an area's temporal components, which allows for nonlinear area-specific deviations. Three scenarios were generated, with 20 datasets simulated under each scenario to assess our proposed model's ability to cluster areas together based on their trends and rates.

The simulation scenarios are defined as follows:

- Simulation data set-up 1: Data generated with change point trend as shown in Figure 5.2 (i).
- Simulation data set-up 2: Data generated with increase and decrease trend as shown in Figure 5.2 (ii).
- Simulation data set-up 3: Data generated with less separated trend as shown in Figure 5.2 (iii).

This chapter develops a spatial-temporal approach and explores spatial-temporal data over a long period of time, including 60 time points from August 2020 to October 2021. This long time frame allowed for a comprehensive analysis of the progression of COVID-19 over time and across different geographical regions; as the real data has a long period, we also used a long period for the simulated data.

Figure 5.3 illustrates plots of the means for the three sets of simulated data.

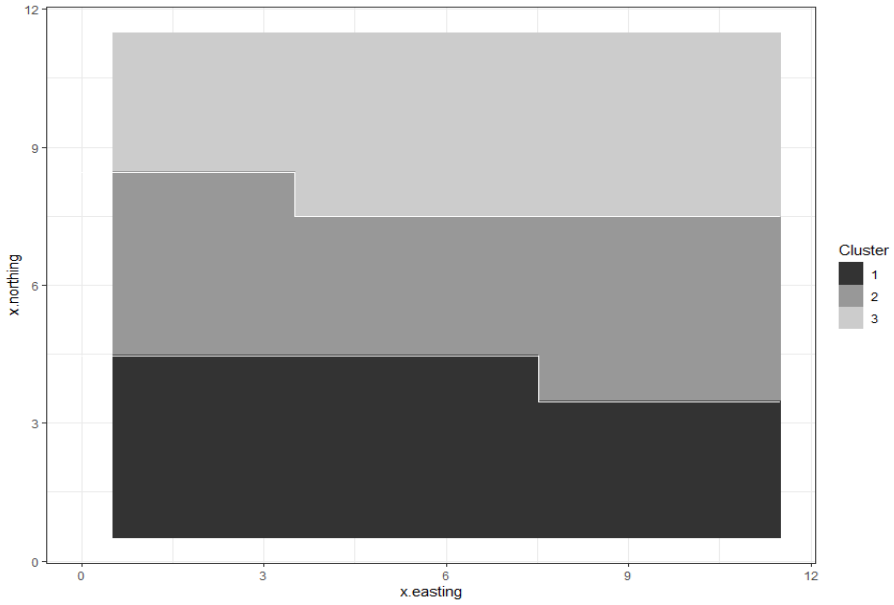


Figure 5.1: Plot of the true three cluster structure where adjacent cell areas tend to have similar distributions

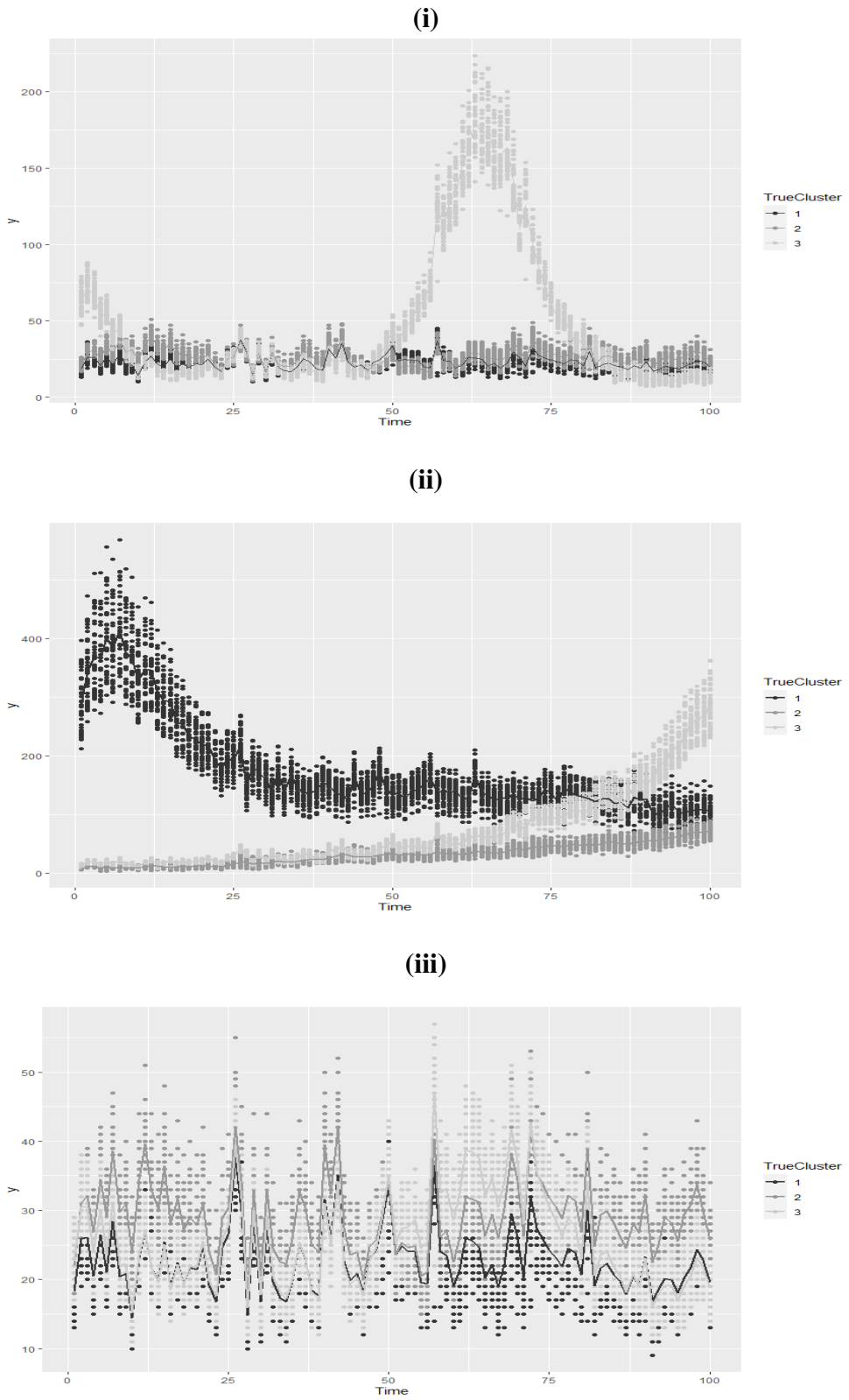


Figure 5.2: Plots of three simulated data sets. (i): Scenario 1 . (ii): Scenario 2. (iii): Scenario 3. The lines represent the mean of each cluster.

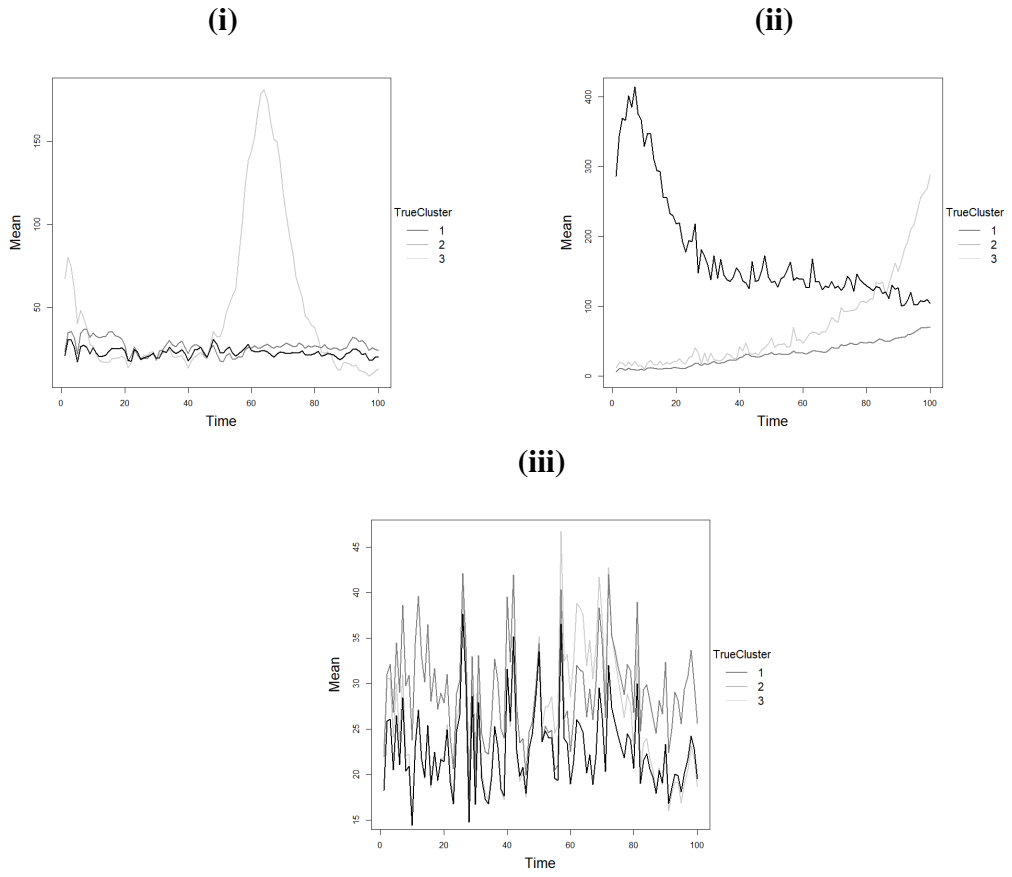


Figure 5.3: Plots of the means of the three simulated data sets. (i): Scenario 1: Data generated with change point trend . (ii): Scenario 2: Data generated with increase and decrease trend. (iii): Scenario 3: Data generated with less separated trend.

5.3.3 Results of the Simulated Study

The study results are presented in Table 5.1, which compares the proposed model with the Bayesian space–time model for clustering areas based on their disease trends using different measures. For the latter, the inference is based on 4,000 samples obtained by generating 100,000 samples, with the first 60,000 discarded as burn in and the remaining 40,000 thinned by 10 to reduce the autocorrelation. The convergence is determined visually by examining the trace plots, which should look weakly stationary. The correctness of the estimated cluster structures is measured by the number of clusters found and the adjusted Rand Index between the true and estimated cluster structures (Hubert & Arabie, 1985). The adjusted Rand index takes a value between 0 and 1, where 1 indicates perfect agreement between the two cluster structures, whereas 0 indicates that the two cluster structures do not agree (see Section 2.13 for more details). The accuracy of the risk surfaces estimated by both models is measured by their root mean square error, $RMSE = \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{y}_{it} - y_{it})^2}$ (Section 2.8.4). We also included the median run time of each model.

From Table 5.1, the median Rand index across 20 replications for our proposed model is higher than the Rand index of the Bayesian space–time model for clustering areas based on their disease trends (STCARclustrends) for all scenarios. These findings attributed to the latter model identifying groups with the same temporal trends, regardless of the rates of those trends (Figure 5.6); the STCARclustrends model aims to find a clustering paradigm to identify groups of areas exhibiting similar temporal trends. As shown in the bottom panel of Figure 5.6, the clusters identified using the STCARclustrends model combine all areas with an increasing trend even though they are generated from different clusters. The table also shows that the RMSE is always lower using the proposed model for the three simulated data compared to the Bayesian space–time model. The proposed model produces results in around a minute, whereas the other model

takes hours. This difference in time is down to one being Bayesian and the other not.

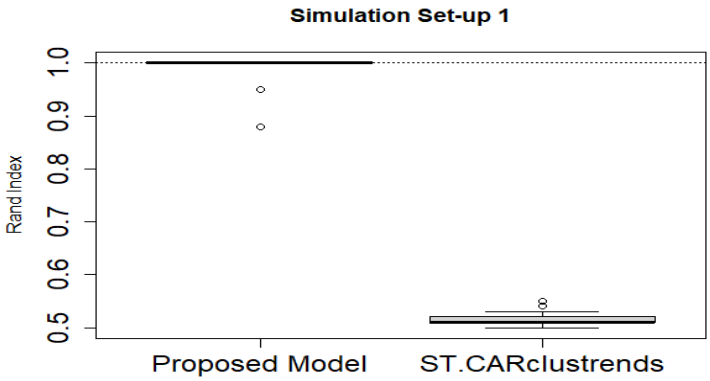
The replicate results are summarised via boxplots, Figure 5.4 displays the adjusted Rand Index values for each model under each simulation set-up. The top panel shows boxplots for simulation set-up 1. Our proposed model performs much better than the STCARclustrends with a value of compared to 0.51. The middle panel shows a boxplot for simulation set-up 2; despite the spread of the boxplot, our model still has a higher Rand index than the other model, with a median of 1 compared to 0.70. The STCARclustrends perform poorly in the bottom panel, which displayed boxplots for simulation set-up 3 with a Rand index as low as 0.22; on the other hand, our model gave a median Rand index equal to 1. This suggests that our model successfully identified, on average, the correct data structure, which was fairly similar to the true structure.

Table 5.1: Simulation results (median across 20 replications) of the proposed model and the STCARclustrends.

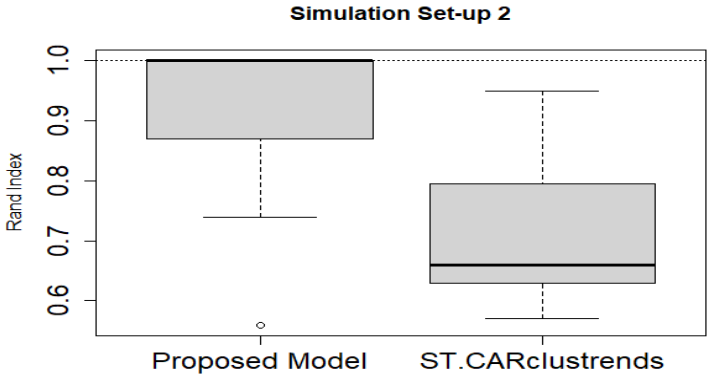
	Simulation Set-up 1	Simulation Set-up 2	Simulation Set-up 3
Proposed Model - No.Cluster	3	3	3
STCARclustrends [a] - No.Cluster	4	3	3
Proposed Model - ARI	1	1	1
STCARclustrends [a] - ARI	0.51	0.70	0.22
Proposed Model - RMSE	6.15	9.16	5.01
STCARclustrends [a] - RMSE	17.60	26.48	5.88
Proposed Model - Time	13.28 mins	8.37 mins	7.49 mins
STCARclustrends [a] - Time	2320 mins	4707 mins	4290 mins

[a] A Bayesian space–time model for clustering areas based on their disease trends

(i)



(ii)



(iii)

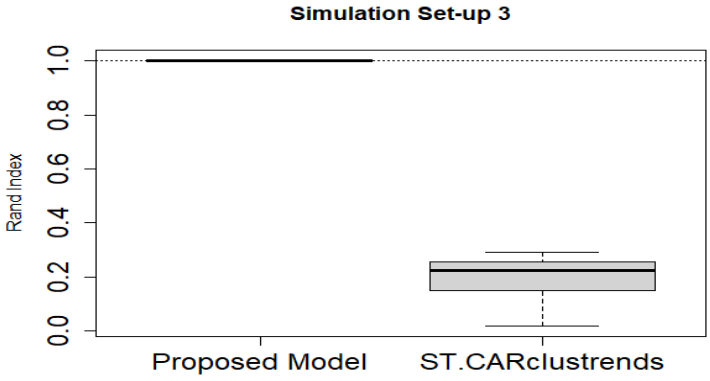
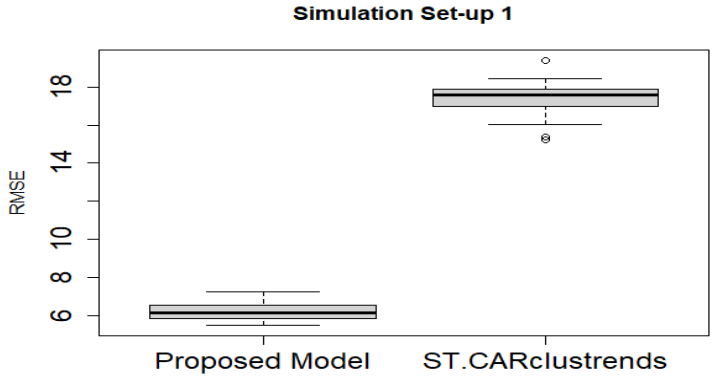


Figure 5.4: Summary of the Rand Index results obtained under each simulation set-up. (i) simulation set-up 1. (ii) simulation set-up 2, and (iii) simulation set-up 3. The dotted lines represent a perfect match between the clustering and the truth.

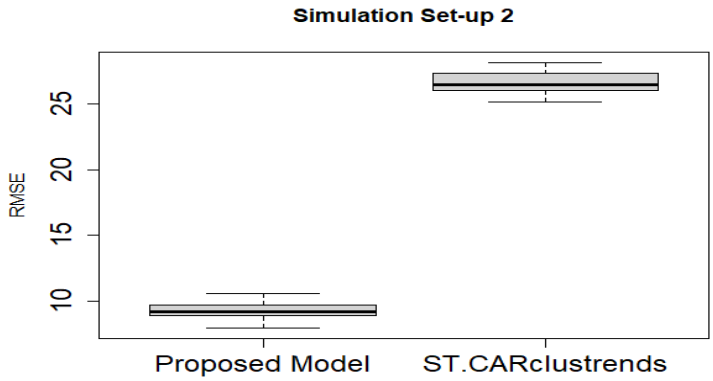
Figure 5.5 presents boxplots showing the estimated data's root mean square error for each approach used in the simulation study. Our proposed model is more precise than the ST.CARclustrends model for the first simulation data (Top panel), since our model has an average RMSE of 6.15 compared to 17.60 obtained from ST.CARclustrends model. Similar findings are archives from simulation set-up 2 (middle panel), which indicate that our model outperforms the ST.CARclustrends model with RMSE equal to 9.16 and 26.48, respectively. For the final simulation set-up (bottom panel), the ST.CARclustrends obtained RMSE equal to 5.88, which is close to 5.01, obtained by our model.

In the case of simulation set-up 2, we plotted the cluster configuration of both approaches for randomly selected simulated data to look for the differences in the classifications. Figure 5.6 presents three panels; the top panel displays simulated data from simulation set-up 2 with the three true clusters, the middle panel shows the fitted values and the estimated clusters obtained from our proposed model, and the bottom panel shows the fitted values and the estimated clusters obtained from ST.CARclustrends model for the simulated data on the top panel. Our proposed model appears to be effective in classifying points within all clusters. On the other hand, the clustering configuration plot of the ST.CARclustrends model (bottom panel) shows how the model finds clusters depending on the trend types being constant and monotonically increasing.

(i)



(ii)



(iii)

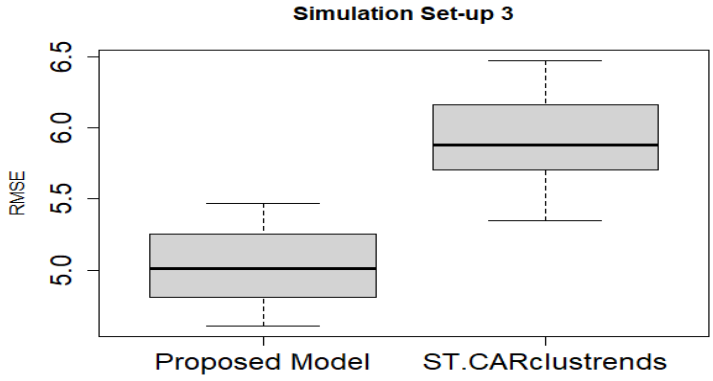


Figure 5.5: Summary of RMSE for the estimated data obtained under each simulation set-up. (i) simulation set-up 1, (ii) simulation set-up 2, and (iii) simulation set-up 3.

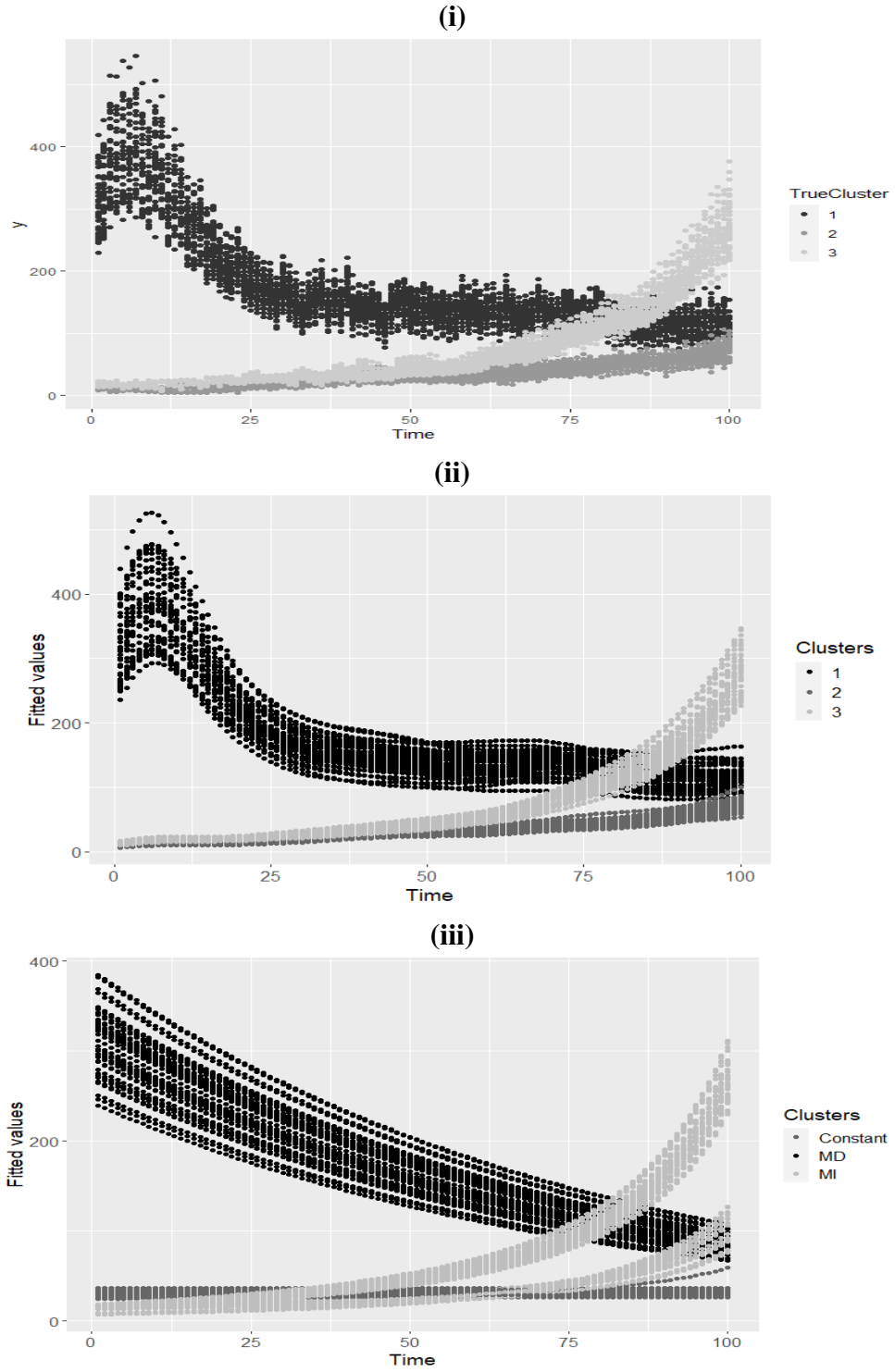


Figure 5.6: (i) displays simulated data from simulation set-up 2 with the three true clusters. (ii) displays the fitted values with three clusters using our proposed model for the simulated data on the top panel. (iii) displays the fitted values with three clusters using the STCARclustrends model for the simulated data in (i).

5.4 Application to Covid-19 data from Glasgow

This section applies our proposed approach to the Greater Glasgow and Clyde Coronavirus disease (COVID-19) data to assess the impact of the worldwide pandemic on this region (Health & Data, 2020). The study area is the Greater Glasgow and Clyde Health Board area, which includes the river Clyde estuary in the west and the city of Glasgow in the east, the largest city in Scotland. The health board of this region is split into $n = 257$ administrative regions known as Intermediate Zones (IZs) as displayed in Figure 3.1. The response data, $\mathbf{y} = (\mathbf{y_1}..., \mathbf{y_n})$, are the 7-day counts of COVID-19 cases from August 2020 to October 2021 (shown in Figure 5.7), with a total of 60-time points for each of the 257 areas. $\mathbf{y_i} = (y_{i1}, ...y_{it})$ and y_{it} represents the 7-day counts of COVID-19 cases in area i at time t, where $(i = 1, .., n)$ and $(t = 1, ..., T)$. Figure 5.7) shows different waves of COVID-19 cases, with the highest peak in September 2021. Most likely, the cause of these waves is the restrictions and rules applied by the governments. The spatial pattern of the 7-day COVID-19 cases for the last week of the study period (18-10-2021) is spatially displayed in Figure 5.8.

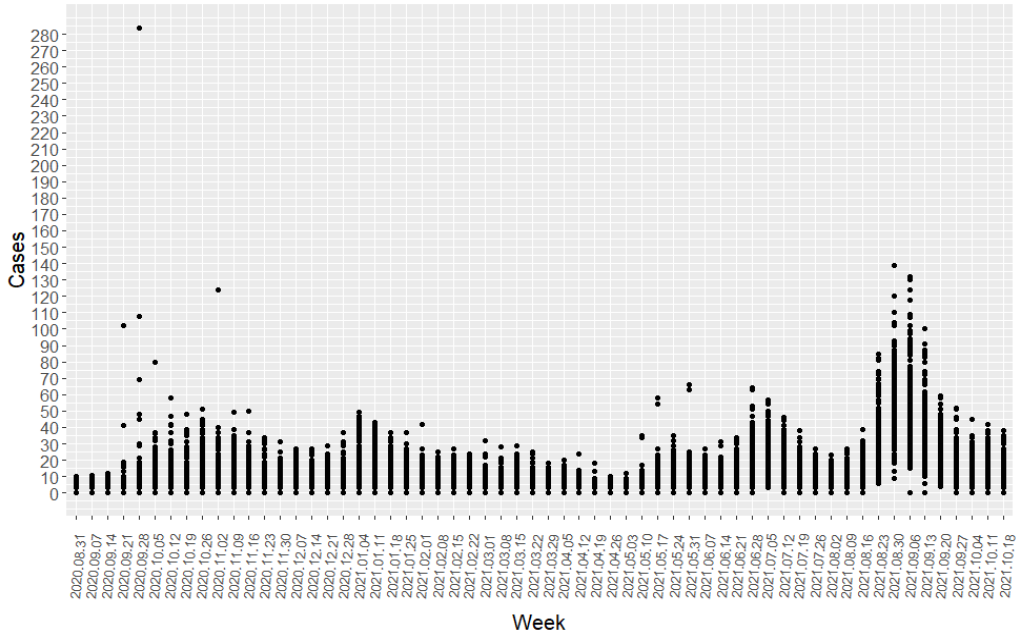


Figure 5.7: A plot of 7-day COVID-19 cases during the study period (August 2020 to October 2021).

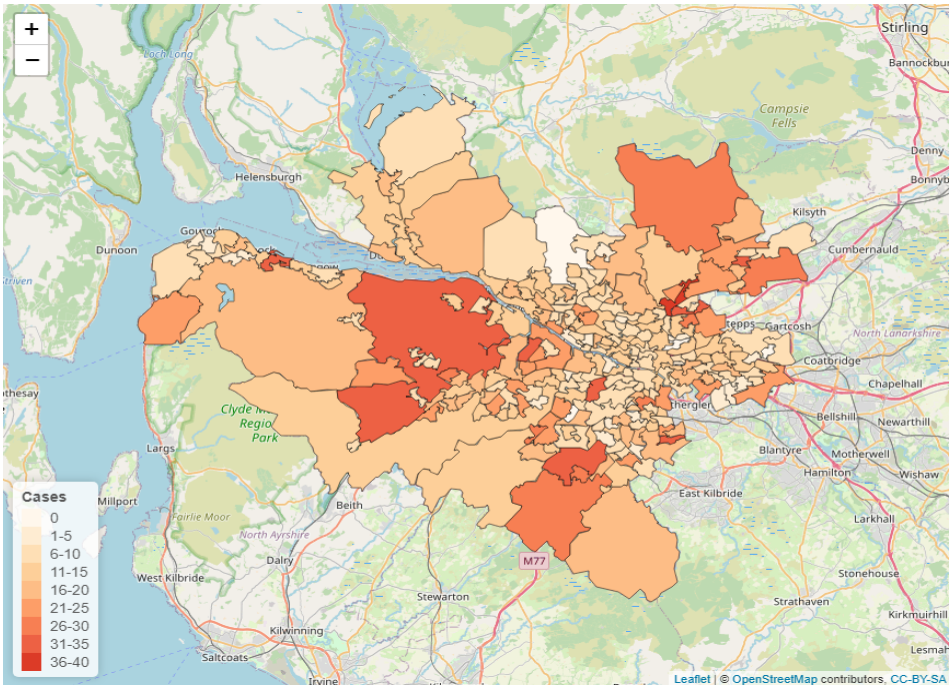


Figure 5.8: A map of Greater Glasgow and Clyde for 7-day COVID-19 cases of the last week of the study period (18-10-2021).

5.4.1 Results

We fit the proposed model described in Section 5.2.1 to the COVID-19 data with different basis dimensions from 10 to 25. In terms of choosing the optimal number of basis dimensions, we check the RMSE. Figure 5.9 displays the RMSE for each number of basis dimensions, which suggests that the RMSE started to level off after 20. therefore, We set our basis dimensions for fixed and random effects to 20.

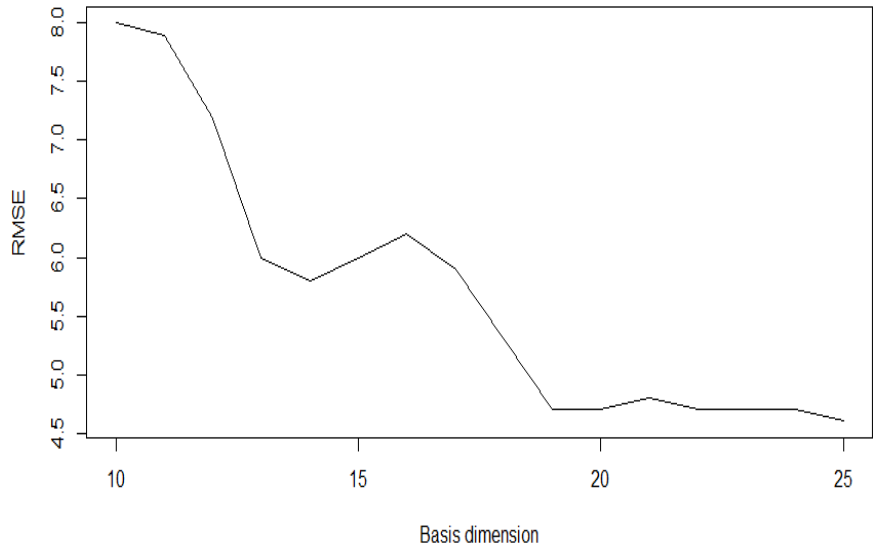


Figure 5.9: Root mean square error (RMSE) for various basis dimensions sizes.

The model identified two clusters for the Greater Glasgow and Clyde Covid-19 cases for our study period via BIC, which is used as the criterion for cluster model selection in `mclust` (Section 2.8.2). Figure 5.10 displays the time series of weekly fitted cases according to clusters obtained from the proposed model. The light colour displays cluster 1, and the dark colour displays cluster 2. The mean line of the two clusters is presented in Figure 5.11, where the solid light line is for cluster 1 with lower COVID-19 cases, and the dark line represents cluster 2. We can see that the temporal pattern is similar in both clusters, but with the light line (cluster 1) generally being lower. The

similar temporal trends in the study region's areas might be due to the local authorities applying some restrictions rules, such as lockdowns, over all of the study region when the number of COVID-19 cases rises to contain the spread of the virus. It has been noticed that areas in both clusters have waves of COVID-19 cases during the study period, with an increase in COVID-19 cases at the end of October 2020 and the beginning of January, July and September 2021.

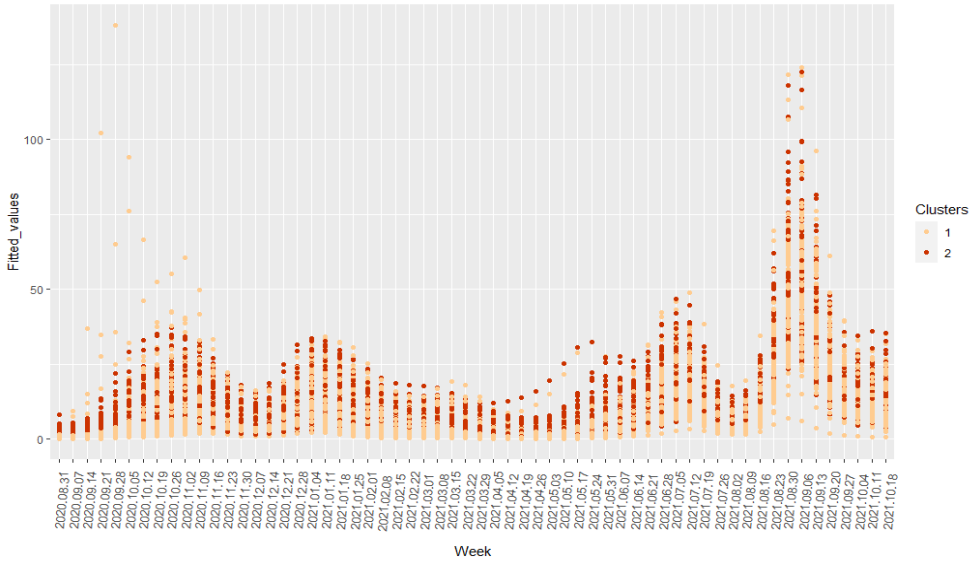


Figure 5.10: Time series plots of the weekly fitted COVID-19 cases in the Greater Glasgow and Clyde from August 2020 to October 2021 with 60-time points. The light colour represents cluster 1, and the dark colour represents cluster 2.

Figure 5.12 displays a map of Greater Glasgow and Clyde, with the two clusters identified by our model, where the dark colour corresponds to cluster 2 with slightly higher COVID-19 cases than cluster 1, which is represented by the light colour. There are 90 areal units in cluster 2 and 167 areal units in cluster 1. The change in the COVID-19 cases is quite smooth across the map, which indicates the presence of spatial autocorrelation within the data. The high-risk cluster mainly included the Renfrewshire, East Renfrewshire, and Glasgow city areas. We present in Figure 5.13 a visual comparison of the actual number of COVID-19 cases in mid-October 2021 (top map), along with

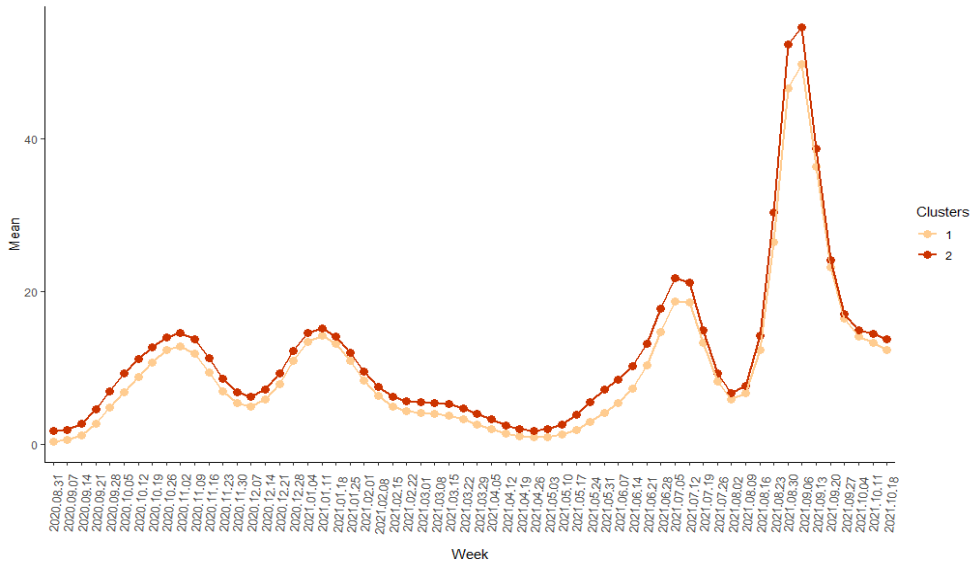


Figure 5.11: Plot of the mean of the number of cases across all areas in clusters obtained from the proposal model where the light colour represents cluster 1, and the dark colour represents cluster 2.

the fitted values of the same date obtained using our model (bottom map). The darker shade corresponds to areas with a high number of COVID-19 cases, whereas the lighter shade corresponds to areas with fewer COVID-19 cases. The maps and the scatter plot (Figure 5.14) show similarity between observed and model fitted values, suggesting that our model has captured the underlying patterns in the COVID-19 data.

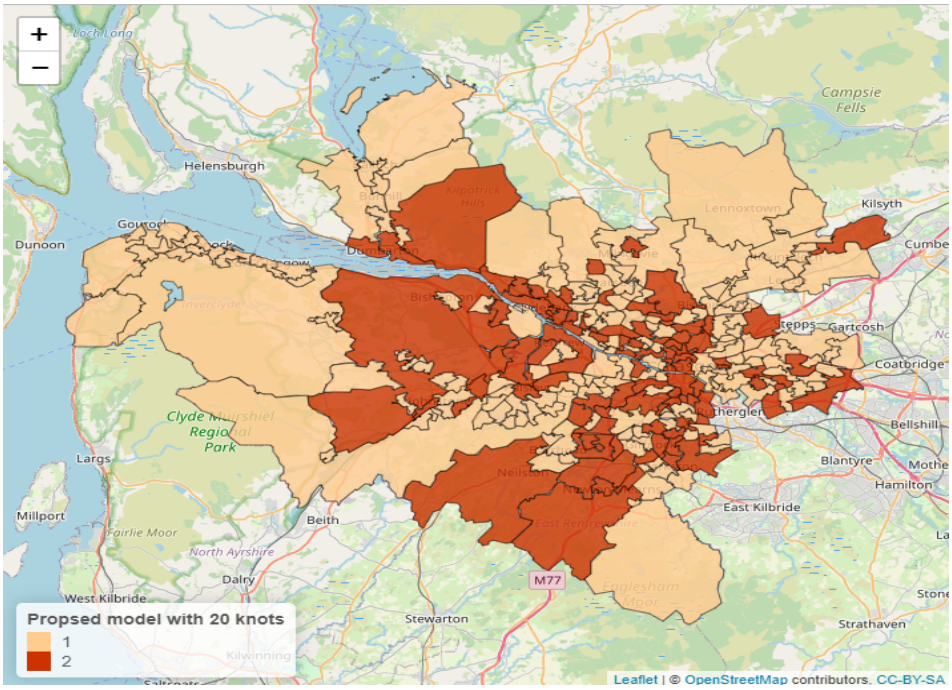


Figure 5.12: A map of Greater Glasgow and Clyde with the estimated clusters. Cluster 2 has higher COVID-19 cases than cluster 1.

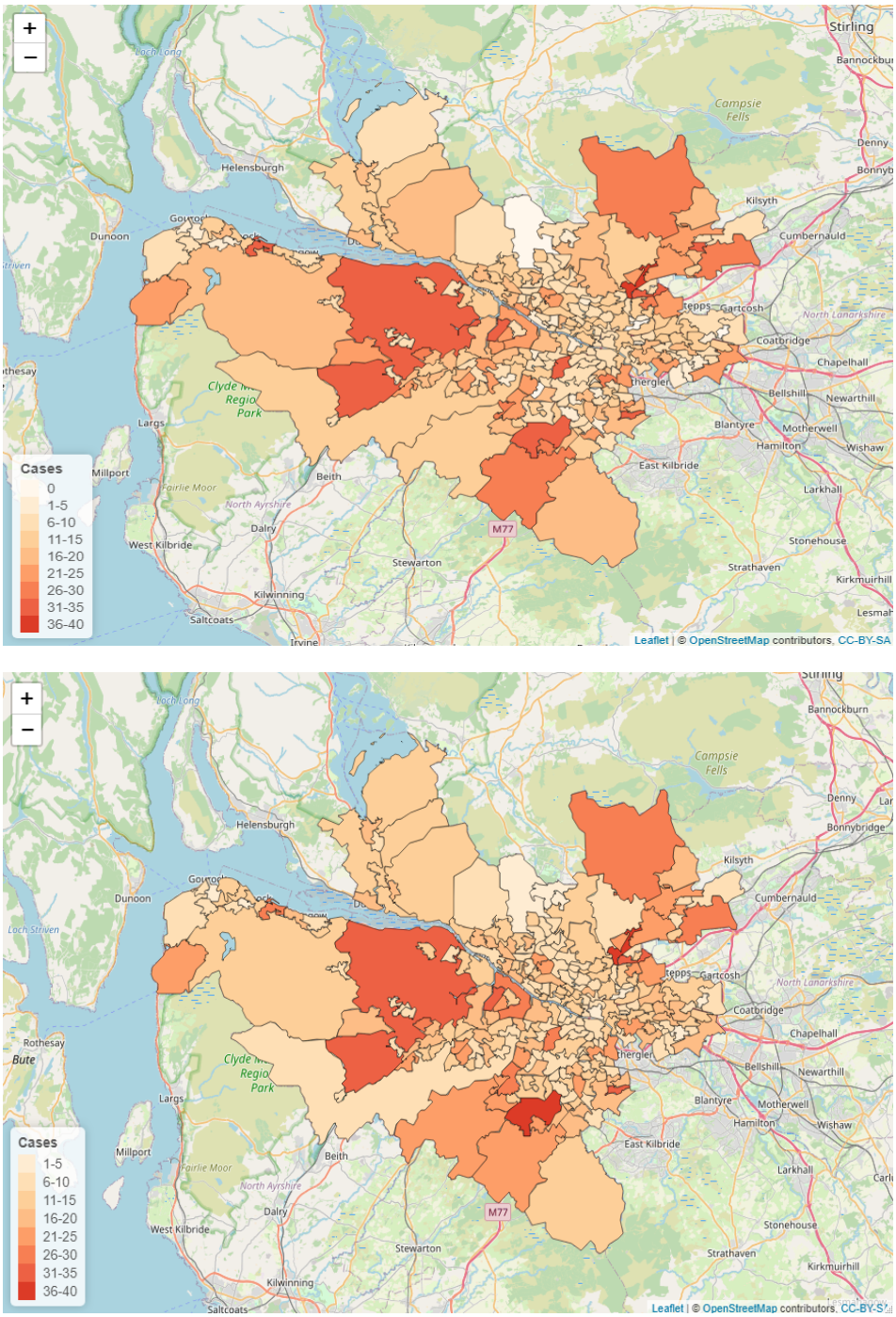


Figure 5.13: The top map displays Greater Glasgow and Clyde with 7-day COVID-19 cases of the last week of the study period (18-10-2021). The bottom map displays Greater Glasgow and Clyde with the fitted values of 7-day COVID-19 cases of the last week of the study period (18-10-2021) using our model

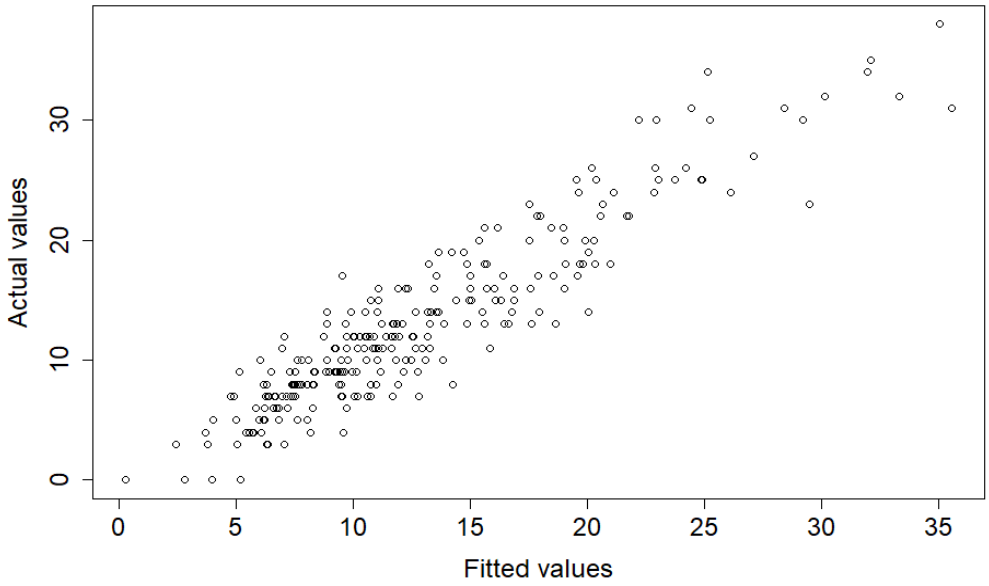


Figure 5.14: A plot of the actual values versus the fitted values of 7-day COVID-19 cases of the last week of the study period (18-10-2021).

5.5 Summary

In this Chapter, we have proposed a two-step spatial-temporal approach to estimate the disease risk pattern and identify clusters of the areas that share similar characteristics in terms of disease risk over our study period. First, a generalised additive mixed model is fitted to the spatio-temporal data inspired by the MacNab & Dean (2001) model. Then, we apply a model-based clustering model (Fraley & Raftery, 2002) to the coefficients of the smoothing term of the space-time interaction to estimate a cluster structure under the model. Note that the clusters in this Chapter are not forced to be spatially contiguous. In Section 5.3, we presented the simulation study results that show our model performs better than the other model in terms of the root mean square error and the Rand Index. For the real data application, the proposed model was applied to the COVID-19 data as shown in Section 5.4 to understand the pattern in the data, which is crucial for public health measurement and resource allocation. The model identifies

two clusters for the study region's areas where both have similar temporal trends with slightly different means, with a peak in COVID-19 case numbers observed in both clusters during the month of September 2021. The slight difference in means suggests limited variability in COVID-19 cases between the identified clusters and the similar temporal trend suggests that the pandemic's temporal dynamics, such as the timing and magnitude of peaks and troughs, are consistent across the study region. The cause of this similarity is possibly due to the country's rules for dealing with the COVID-19 epidemic, which are implemented in all areas at the same time, such as lockdowns, mask usage, and social distance guidelines.

Chapter 6

Spatio-temporal model with a cluster factor

Chapter 4 aimed to detect groups of areal units with similar disease risk at a specific time point, including the spatial information. This approach incorporated the spatial information of the areal data via a Gibbs prior. Chapter 5 introduced a two-step modelling approach for estimating and detecting clusters for spatio-temporal data. In the first step, a generalised additive mixed model was fitted to the spatio-temporal data and then applied a model-based clustering model to the coefficients of the smoothing terms of the space-time interaction to estimate a cluster structure under the model.

Given the potential clustering structure in spatial data over time, adding a cluster factor to a spatio-temporal model might improve the estimation of disease risk across study areas. In this chapter, an extension of the spatio-temporal model which was introduced in the previous chapter is presented. This will include a cluster membership factor as a new term in the previous model, since it is believed that there is a potential clustering structure in spatial data over time.

In this chapter, Section 6.1 presents the proposed approach and the theoretical and com-

putational framework. Following this, Section 6.2 presents the simulated data results, where we display the outcome derived from our approach and compare it to another spatio-temporal model. In Section 6.3, we apply the proposed approach to the Coronavirus disease (COVID-19) count data for the Greater Glasgow and Clyde Health Board during the time period from August 2020 to October 2021. Lastly, Section 6.4 provides an overview of this approach's essential findings and benefits.

6.1 Methodology

This section proposes an extension to the spatio-temporal generalised additive mixed model introduced in Chapter 5. This approach is a two-step approach for identifying the clusters and estimating spatio-temporal disease risk of the spatio-temporal data. In the first step, apply the model-based clustering approach introduced in Section 2.11.1 to the multivariate SIR disease risk of the response variable of each area over the study time period for the spatio-temporal data. In the second step, the result of the clustering approach is added as a factor to the model introduced in Chapter 5, the spatio-temporal generalised additive mixed model, to estimate the disease risk.

6.1.1 Proposed model

We propose an extended spatio-temporal generalised additive mixed model for count data outcomes that includes a cluster factor, spatially correlated random effects via the conditional autoregressive (CAR) model (Section 2.9.2), fixed effects P-splines smoothing for modelling the overall temporal trends, and random effects P-splines for modelling space–time (Section 2.7.2).

Let y_{it} be the observed number of cases for the i^{th} area at the time t, where $i = 1, 2, ..., n$, and $t = 1, 2, ..., T$. Suppose the data are clustered into g clusters where $I(\text{cluster}_i = j)$

is 1 if area i is in cluster j, 0 otherwise and $j = 1, ..., g$.

The general formula will be as follows:

$$y_{it} \sim \text{Poisson}(\mu_{it})$$
$$\log \mu_{it} = \log n_{it} + \log \varpi + \sum_{j=2}^{g} \zeta_j I(\text{cluster}_i = j) + A(t) + \phi_i + \beta_i(t) \tag{6.1.1}$$

where n_{it} represents the population in the i^{th} area at the time t, ϖ is the mean rate over all times and areas, ζ_j is the cluster-specific intercept, and ϕ_i is the spatial random effect. $A(t)$ is the fixed effects smoothing function of t, and $\beta_i(t)$ is the random effects smoothing function of the space–time interaction.

After applying a P-spline for the fixed and the random temporal effect, the model will be defined as follows:

$$\log \mu_{it} = \log n_{it} + \log \varpi + \sum_{j=2}^{g} \zeta_j I(\text{cluster}_i = j) + S_0(t) + \phi_i + S_i(t) \tag{6.1.2}$$

where ζ_j is the cluster-specific intercept, ϕ_i modelled using an intrinsic CAR model, $\boldsymbol{\phi} = (\phi_1, ..., \phi_i) \sim N(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W})^-)$, where $\mathbf{Q}(\mathbf{W})$ is a precision matrix, $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$, $\mathbf{W1}$ is a vector containing the number of neighbours for each areal units. $\mathbf{W}$ is a neighbourhood matrix, and τ is a scalar precision parameter. The smooth function $S_0(t) = \sum_{k=1}^{K} p_k(t)\beta_{0,k}$ is a set of P-splines for modelling the overall temporal trend, and $S_i(t) = \sum_{k=1}^{K} p_k(t)B_{i,k}$ is a P-spline smoothing for the time components of area i which allow for nonlinear area-specific deviations. The number of the P-spline basis functions is denoted by K and p_k, $k = 1, ..., K$, represents the basis function. The fixed effect vector is $\mathbf{a} = (\log \varpi, \zeta_j, \beta_{0,1}, ..., \beta_{0,K})$, whereas the random effect is $\mathbf{b} = (\phi_1, ..., \phi_i, ..., \phi_n, B_{1,1}, ..., B_{1,K}, ..., B_{i,k}, ..., B_{n,K})$.

6.1.2 Estimating the model

Step 1: Model-based Clustering:

In the first step, the model-based clustering approach, introduced in Section 2.11.1, is applied to the rate of data $\mathbf{R}$, which is the observed value of counts divided by the population sizes for each time point i.e. $\mathbf{R_i} = (R_{i1}, ..., R_{iT})$ where $R_{it} = y_{it}/n_{it}$. The finite mixture model with g groups is defined by

$$f(\mathbf{R}_i) = \sum_{j=1}^{g} \pi_j f_j(\mathbf{R}_i|\mu_j, \Sigma_j) \tag{6.1.3}$$

The likelihood function is:

$$L(\mu, \Sigma, \pi|\mathbf{R}) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j f_j(\mathbf{R}_i|\mu_j, \Sigma_j) \tag{6.1.4}$$

The expectation maximization (EM) algorithm (Dempster et al., 1977) is used for estimating mixture model parameters. In this setting, we consider the complete dataset to be viewed as $(\mathbf{R}_i, \mathbf{z}_i)$ where

$$z_{ij} = \begin{cases} 1 & \text{if } \mathbf{R}_i \text{ belongs to component } j, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the likelihood of the complete data will be defined as

$$L_{\text{complete}}(\mu, \Sigma, \pi, \mathbf{z}|\mathbf{R}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\pi_j f_j(\mathbf{R}_i|\mu_j, \Sigma_j)\right)^{z_{ij}}, \tag{6.1.5}$$

where:

- π_j is the prior probability of membership of group j, where $j = 1, ..., g$.
- $f_j(\mathbf{R}_i|\mu_j, \Sigma_j)$ is the density of a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j.
- z_{ij} is a binary indicator variable that equals 1 if the observation i is belong to cluster j, and 0 otherwise.

For $v = 2, 3...$ repeat the E and M steps in turn until convergence is reached.

E-Step (Expectation)

$$\hat{z}_{ij}^{(v)} = \frac{\hat{\pi}_j^{(v-1)} f_j(\mathbf{R}_i|\hat{\mu}_j^{(v-1)}, \hat{\Sigma}_j^{(v-1)})}{\sum_{l=1}^{g} \hat{\pi}_l^{(v-1)} f_l(\mathbf{R}_i|\hat{\mu}_l^{(v-1)}, \hat{\Sigma}_l^{(v-1)})}$$

M-Step (Maximization)

$$n_j^{(v)} = \sum_{i=1}^{n} \hat{z}_{ij}^{(v)}$$

$$\hat{\pi}_j^{(v)} = \frac{n_j^{(v)}}{n}$$

$$\hat{\mu}_j^{(v)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(v)} \mathbf{R}_i}{n_j^{(v)}}$$

$$\hat{\Sigma}_j^{(v)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(v)} \left(\mathbf{R}_i - \hat{\mu}_j^{(v)}\right) \left(\mathbf{R}_i - \hat{\mu}_j^{(v)}\right)^{\top}}{n_j^{(v)}}$$

More details for calculating $\hat{\Sigma}_j^{(v)}$ under different restrictions are given by Celeux & Govaert (1995) and Scrucca et al. (2016). Within the `mclust` framework, the optimal number of clusters is selected using the negative Bayesian information criterion (Section 2.8.2), where the number of clusters with the highest BIC value is preferable.

The posterior probabilities of data points belonging to each cluster are computed by model-based clustering, and each point is assigned to the cluster with maximum a posteriori probability, giving a discrete (hard) cluster assignment variable.

Step 2: The spatio-temporal generalised additive mixed model:

The model in step 2 of this approach is fitted using `mgcv` package as in Section 5.2.2, the coefficients of the model are estimated by penalized maximum likelihood estimation. For equation (6.1.2), $S_0(t)$ estimated using penalized iteratively re-weighted least squares(PIRLS) algorithm. The estimation of the smooth parameter is the restricted maximum likelihood (REML), and $S_i(t)$ by adapting a Bayesian manner with a prior, $B \sim MVN(\mathbf{0}, (\lambda \mathbf{P})^{-1})$.

6.2 Simulation study

In this section, we generated various simulation scenarios to evaluate how well the proposed model introduced in Section 6.1.1 performs. The scenarios include similar temporal trends with various clusters' intercepts. A comparison was conducted between the proposed model, the generalised additive mixed model with clusters (GAMM with clusters), and the generalised additive mixed model without clusters (GAMM without clusters). The results of this simulation study are outlined below.

6.2.1 Data Generation

This chapter generated simulated spatial data by first creating a grid. The data generation process involved fitting a generalized additive mixed model (GAMM) and incorporating cluster intercepts based on their positions as neighbours within the grid. The total number of areas within each cluster was chosen to achieve approximate equality across all clusters, ensuring balanced representation. The simulated data were generated for $n = 121$ areas and $T = 30$ time points using a generalized additive mixed model with Poisson distribution with three clusters.

The data were generated using the following model:

$$\begin{aligned} y_{it} &\sim \text{Poisson}(\mu_{it}) \quad i = 1,...,121, \quad t = 1,...,100, \\ \log \mu_{it} &= \sum_{j=1}^{g} \zeta_j I(\text{cluster}_i = j) + S_0(t) + \phi_i + S_i(t) \quad j = 1,2,3, \\ \phi_i &\sim N(\mathbf{0}, Q^{-1}). \end{aligned} \tag{6.2.1}$$

We generated cubic P-splines for the temporal trend $S_0(t)$ (outlined in detail in Section 2.7.2). We added random effects generated from a multivariate Gaussian distribution with a spatially correlated precision matrix, $Q = \text{diag}(W1) - W$, where W is a neighbourhood matrix, which is specified as $w_{ij} = 1$ if the two areas are sharing a common border and $w_{ij} = 0$ if they do not share a common border. The intrinsic CAR (ICAR) model, which is defined in Section 2.9.2, was used for the spatial effects (Besag et al., 1991). $S_i(t)$ is a P-spline smoothing for an area's temporal components, which allows for nonlinear area-specific deviations. The clusters intercept $\zeta = (\zeta_1, \zeta_2, \zeta_3)$ is included with different scenarios as defined below.

Three scenarios were generated, with 50 replicated datasets simulated under each scenario to assess our proposed model's ability to estimate disease risk.

Scenarios

- Set-up 1

  Three clusters with a small difference $\zeta = (1.5, 1, 0.5)$ as shown in Figure 6.1 (i).

- Set-up 2

  Three clusters with a large difference $\zeta = (1, 0, -1)$as shown in Figure 6.1 (ii).

- Set-up 3

  Three clusters with moderate difference $\zeta = (2, 1.2, 0)$ as shown in Figure 6.1 (iii).

This chapter focused on COVID-19 data during periods of fewer restrictions, where the proposed approach was applied to the period from March 2021 through October 2021. Since the number of the real data were lower in this chapter, the generated data were created to mimic this. The number of replicated datasets were increased as in this chapter both the proposed model and the comparative model rely on frequentist statistics rather than Bayesian methods, so the computational cost is not an issue here.

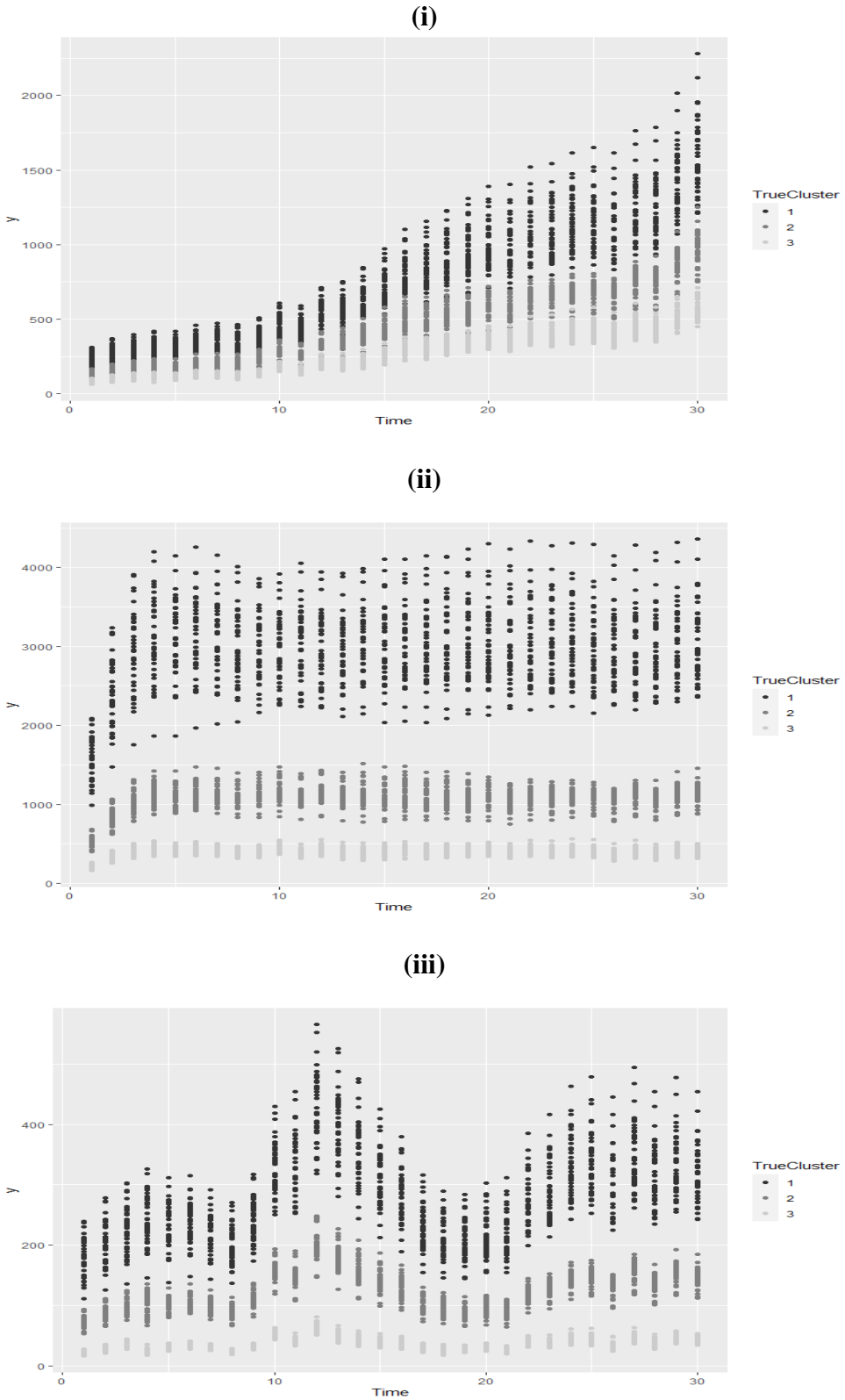


Figure 6.1: Plots of three simulated data sets. (i): Scenario 1 . (ii): Scenario 2. (iii): Scenario 3.

6.2.2 Results of the Simulated Study

The study results are presented in Table 6.1, which compares the proposed model of the spatio-temporal generalised additive mixed model with a cluster factor with the generalised additive mixed model without a cluster factor. The accuracy of the risk surfaces estimated by both models is measured by their root mean square error, $RMSE = \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{y}_{it} - y_{it})^2}$ (Section 2.8.4), whereas the Akaike information criterion (AIC) (Section 2.8.1) is used for model selection, which balances between the model fits and the model complexity. A model with a lower AIC value is often considered preferable.

From Table 6.1, the AIC medians for our proposed model (GAMM with clusters) across 50 replications for each of the three simulated datasets are 32331, 36191, and 27783. On the other hand, the AIC median values for the model without the clusters for the same three simulated datasets are 32333, 36192, and 27791. Our proposed model obtains a lower AIC median than the GAMM without clusters for all simulated data, indicating that our model is preferable and a better fit. Both models, GAMM with clusters and GAMM without clusters, exhibit similar root mean square error (RMSE) median values across 50 replications for each simulated dataset with RMSE values equal to 19.38, 33.19, and 12.20, respectively. GAMMs with and without a cluster factor have similar accuracy in disease risk estimation, and the difference in the AIC values of both models is quite small. However, the slight preference for our proposed model based on AIC values suggests that our proposed model, which incorporates a cluster factor, may offer some advantages in capturing underlying patterns and essential information in the datasets in comparison to the GAMM without a cluster factor. Furthermore, The results show that the p-values associated with the cluster factor were statistically significant, at the 5% level. This suggests strong evidence against the null hypothesis (no difference between groups) and indicates that the cluster membership factor affects the response

variable in both clusters. The cluster membership factor is likely an important predictor of the response variable in our model, and its inclusion helps explain variation in the response within and between clusters.

Table 6.1: Simulation results (median across 50 replications) of the proposed model (GAMM with clusters) and the model without cluster (GAMM without clusters)

	Simulation Set-up 1	Simulation Set-up 2	Simulation Set-up 3
Proposed Model - AIC	32331	36191	27783
GAMM without clusters - AIC	32333	36192	27791
Proposed Model - RMSE	19.38	33.19	12.20
GAMM without clusters - RMSE	19.38	33.19	12.20

6.3 Application to COVID-19 data

This section applies our proposed approach to Greater Glasgow and Clyde Coronavirus disease (COVID-19) data, which is presented in Chapter 3. As in the previous chapter, the study area is the Greater Glasgow and Clyde Health Board area, and the health board of this region is split into $n = 257$ administrative regions known as Intermediate Zones (IZs) as displayed in Figure 3.1. The response data, $\mathbf{y} = (\mathbf{y_1}..., \mathbf{y_n})$, are the 7-day counts of COVID-19 cases from the end of March 2021 to October 2021 (Figure 6.2), with a total of 30-time points for each of the 257 areas. $\mathbf{y_i} = (y_{i1}, ...y_{it})$ and y_{it} represents the 7-day counts of COVID-19 cases in area i at time t, where $(i = 1, .., n)$ and $(t = 1, ..., T)$. The plot shows different waves of the cases, with the highest peak in September 2021. Most likely, the reason for these waves is the government's restrictions and rules.

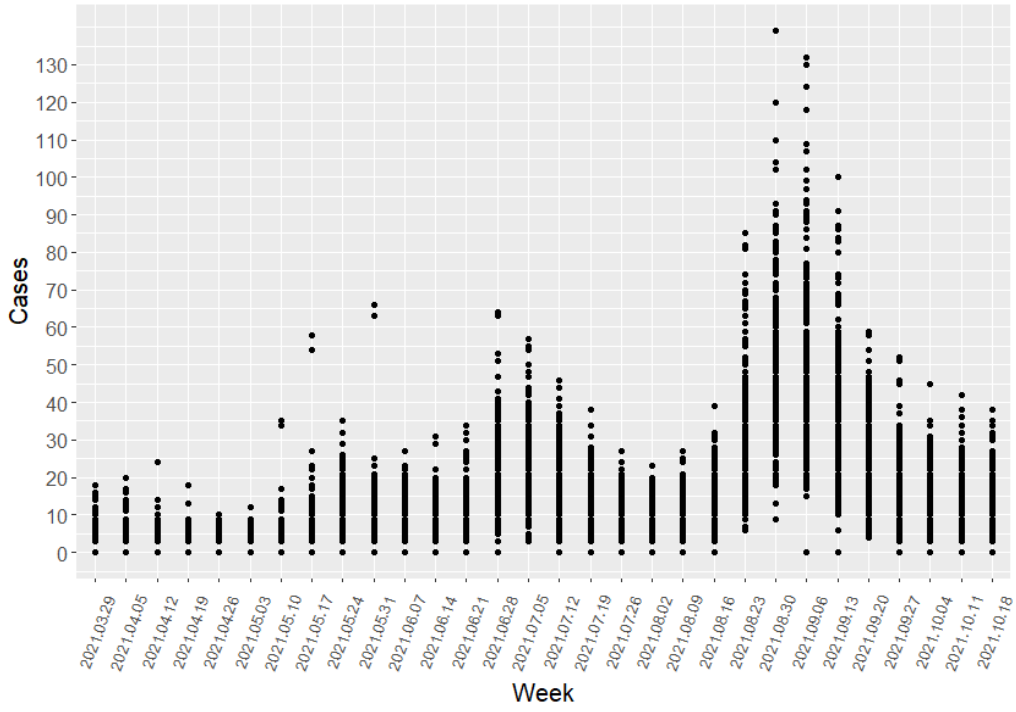


Figure 6.2: A plot of 7-day COVID-19 cases during the study period (The end of March 2021 to October 2021).

6.3.1 Results

The proposed model described in Section 6.1.1 fits the COVID-19 data with different basis dimensions from 5 to 20. To choose the optimal number of basis dimensions, we check the RMSE. Figure 6.3 displays the RMSE for each number of basis dimensions, which suggests that the RMSE started to level off after 15. As a result, We set our basis dimensions for fixed and random effects to 15.

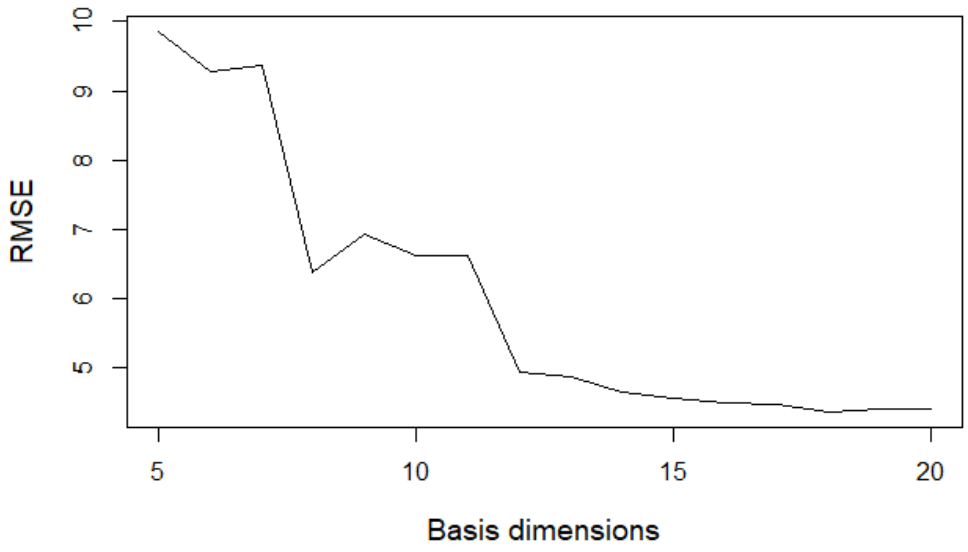


Figure 6.3: Root mean square error (RMSE) for various basis dimensions sizes.

The model identifies two clusters with evidence suggesting that the clusters are statistically significant. The resulting p-values provide evidence against the null hypothesis of no difference in COVID-19 cases across the clusters. This suggests that there are statistically significant differences in COVID-19 cases across the clusters, and these differences are unlikely to have occurred by random chance alone. The cluster factor is likely an essential predictor of the COVID-19 cases in our model, and its inclusion helps explain variation in the COVID-19 cases within and between clusters. Figure

6.4 visually compares the actual number of COVID-19 cases in mid-October 2021 (top map) and the fitted values derived using our model for the same date (bottom map). The darker shade refers to areas with a large number of COVID-19 cases, whereas the lighter shade indicates areas with fewer COVID-19 cases. The change in the COVID-19 cases is relatively smooth throughout the map, indicating the existence of spatial auto-correlation within the dataset. The areas with the highest number of COVID-19 cases are located in East Dunbartonshire, followed by Renfrewshire and East Renfrewshire. The maps and scatter plot (Figure 6.5) reveal that the observed and model-fitted values align closely, indicating strong correspondence between the observed and model-fitted values. The alignment is nearly straight, suggesting our model's effectiveness in catching the underlying patterns in the COVID-19 data.

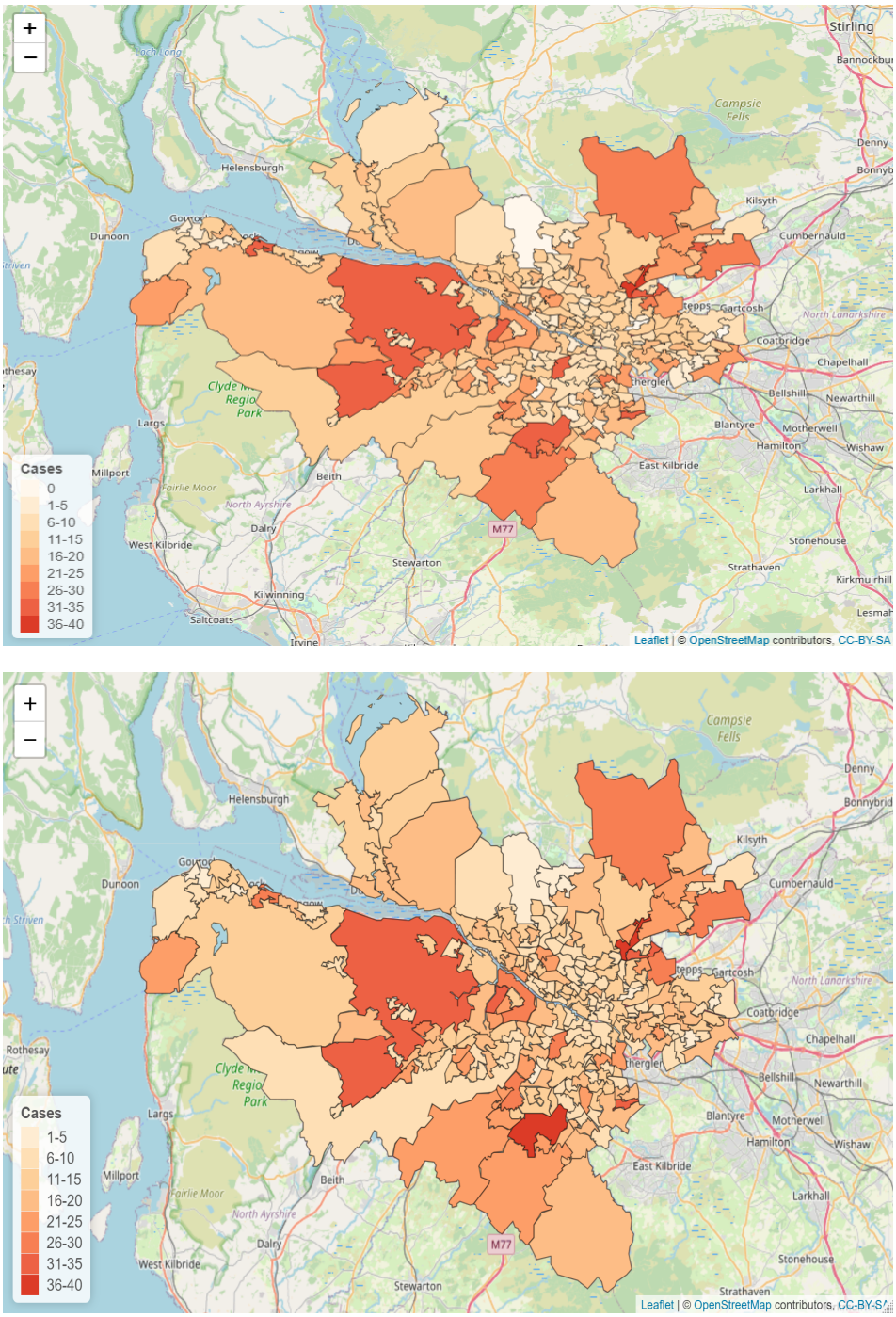


Figure 6.4: The top map displays Greater Glasgow and Clyde with 7-day COVID-19 cases of the last week of the study period (18-10-2021). The bottom map displays Greater Glasgow and Clyde with the fitted values of 7-day COVID-19 cases of the last week of the study period (18-10-2021) using our proposed model

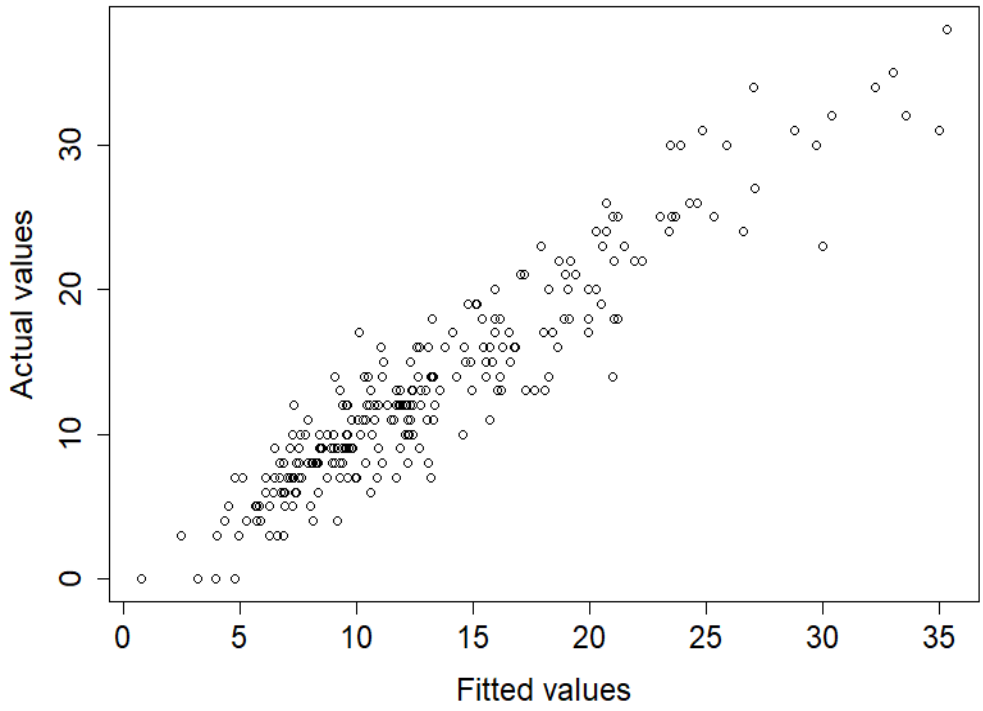


Figure 6.5: A plot of the actual values versus the fitted values of 7-day COVID-19 cases of the last week of the study period (18-10-2021).

6.4 Summary

In this Chapter, we have proposed a spatio-temporal generalised additive mixed model with clusters factor to estimate the disease risk over our study period considering the probable clustering structure in spatial data over time. This approach is a two-step approach for identifying the cluster structure using the model-based clustering (Fraley & Raftery, 2002) and then adding it to the generalised additive mixed model, which is introduced in the first step of the proposed model in Section 5 to estimate the disease risk of the spatio-temporal data. In Section 6.2, we presented the findings of the simulation study, illustrating that adding the clustering factor to a spatio-temporal generalised additive mixed model will demonstrate greater robustness and efficiency in

capturing underlying patterns and essential information in a different dataset. The proposed model fits the COVID-19 data introduced in Section 6.3. The results show that COVID-19 cases exhibit relatively smooth changes across the map, where areas with the highest counts of COVID-19 cases are situated in East Dunbartonshire, with Renfrewshire and East Renfrewshire followed suit. Plotting the scatter plot for observed and fitted values from the proposed model reveals a close alignment that reflects the model's effectiveness in acting the underlying pattern of the data.

Chapter 7

Conclusion

This thesis focused on estimating the disease risk as well as identifying the number and structure of disease risk clusters for spatial and spatio-temporal data. For areal data such as that examined in this thesis, the study region is partitioned into a set of non-overlapping areal units, and each area's disease risk is estimated. Disease mapping helps to visualise the risk patterns by using different colours for areas according to their disease risk. These risks can be estimated using mixed models, which include fixed effects for the overall mean risk level and random effects to control area-specific risks. The primary aim of this thesis was to develop a new approach for spatial and spatio-temporal data to estimate the disease risk and identify disease risk clusters simultaneously.

Chapter 2 introduced the general statistical methodology and inference methods used across this thesis, including frequentist and Bayesian statistics, the conditional autoregressive (CAR) models, and clustering approaches. Chapter 3 gave an introduction to disease mapping and the Coronavirus disease (COVID-19) epidemic data as well as providing a summary of the spatial and spatio-temporal literature. In Chapter 4, a spatially constrained Poisson Finite Mixture model was proposed, which incorporated spatial information into a Poisson Finite Mixture Model via a Gibbs prior, which

allowed regions to be grouped together based on both risk levels and geographical proximity. Chapter 5 introduced a spatio-temporal generalised additive mixed model which fits spatially correlated random effects via the conditional autoregressive (CAR) model, while the P-spline smoothing is used for the fixed and random temporal components to estimate the disease risk over time. Then, a model-based clustering algorithm clusters the P-spline coefficients of the interaction between time and space to identify clusters of the spatio-temporal disease data. In Chapter 6, we considered the potential clustering structure in spatial data over time and presented a two-stage approach which identified the number of clusters and added the structure of these clusters as a factor to the spatio-temporal generalised additive mixed model.

7.1 A Spatially Constrained Poisson Finite Mixture Model

Chapter 4 outlined a new approach for clustering spatial count data using a finite mixture model. The model takes into account the spatial information of the areal data, where areas should belong to the same cluster as most of its neighbours unless there is a high difference in some attributes. The model extends the classical finite mixture model to a spatially constrained Poisson finite mixture model by incorporating spatial information in the individual-specific mixing proportions via a Gibbs density function Markov Random Field-based prior.

Section 4.4.2 outlined the simulation study, which compared this proposed model to the Ward-like hierarchical Clustering model. The results showed that both models identified the correct number of clusters, but our model performs better than the Ward-like hierarchical Clustering model in terms of the Rand Index, where the proposed model had a high Rand index value for the three different simulated data scenarios. In Section 4.5, the proposed model was applied to an application of real data, the Coronavirus

disease (COVID-19) count data for the Greater Glasgow and Clyde Health Board. The model identifies four clusters for the study areas with high-risk areas, including Renfrewshire, East Renfrewshire, and Easterhouse.

7.2 Spatio-temporal modelling and clusters detection

Chapter 5 introduced a spatio-temporal modelling approach for estimating the disease risk over time and identifying the number of clusters for the disease data. The proposed spatio-temporal generalised additive mixed model uses fixed effect smoothing P-splines for the overall temporal effect and random effect P-splines for the interaction between time and space. In addition, the conditional autoregressive (CAR) model is used to estimate the spatial pattern in the data. Fitting this model uncovers the underlying spatial-temporal pattern in the data through the space-time effect. Moreover, the second step in this approach identifies the number of clusters for the spatial-temporal disease data that share a similar trend and rate by applying the model-based clustering algorithm to the estimated P-spline coefficients of the interaction between time and space. The Bayesian information criterion (BIC) is used to select the number of clusters.

A simulation study presented in Section 5.3 showed that our approach performed better than the Bayesian space-time model in terms of identifying the correct number and structure of clusters. In addition, the estimated disease risk derived from our proposed model outperforms the other model. The proposed model was applied to the COVID-19 data as shown in Section 5.4. The model detected two clusters within the areas of the study, both exhibiting similar temporal trends with slightly different means, possibly influenced by the country's rules in managing the COVID-19 epidemic. Renfrewshire, East Renfrewshire, and Glasgow City were all part of the high-risk cluster.

7.3 Spatio-temporal model with a cluster factor

Chapter 6 introduced a variation of the spatio-temporal generalised additive mixed model outlined in Chapter 5. This approach first selects the number and structure of clusters for the dataset by fitting a model-based clustering algorithm on the case rate of the disease in each area (number of cases divided by the population size of each area). The cluster structure obtained from the first step is added as an additional factor in the spatio-temporal generalised additive mixed model, outlined introduced in Chapter 5. This new spatio-temporal generalised additive mixed model included fixed effects terms for the different risk levels that are assigned to each cluster and smoothing P-splines for the overall temporal effect. Also, random effects are included for the interaction between time and space using smoothing P-splines and spatial effects using the CAR model.

A simulation study was carried out to compare our proposed model GAMM with clusters to a standard GAMM without clusters; the proposed model was favoured by AIC, which indicates a better trade-off between model fit and complexity (Section 6.2). The model was then applied to the COVID-19 data for the Greater Glasgow and Clyde Health Board areas, showing that COVID-19 cases appear to be changing relatively smoothly across the map. When the observed and fitted values from our proposed model were plotted together, a close alignment was seen, which indicates that the model successfully captures the underlying pattern of the COVID-19 data (Section 6.3).

7.4 Limitations and future work

This thesis introduces three new approaches. Chapter 4 proposed a new approach to grouping spatial data and estimating the disease risk. This model incorporates spatial dependency while clustering that data. The model was conceptually straightforward to implement via a series of MCMC steps. The drawbacks of these spatially constrained Poisson finite mixture models, particularly when employing Markov Chain Monte Carlo (MCMC) methods is that the MCMC approach is sensitive to the choice of a proposal function. In this thesis, both symmetric and asymmetric settings for the hyperparameters of the Dirichlet proposal function were investigated. In the symmetric configuration, all hyperparameters were assigned identical values, while in the asymmetric configuration, different hyperparameters were permitted. The symmetric setup outperformed the asymmetric one, demonstrating acceptance rates varying from 0.2 to 0.6. The MCMC is computationally intensive and typically requires longer processing times than the Integrated Nested Laplace Approximations (INLA) methodology (Rue et al., 2009), which could be used in future work since INLA is computationally faster than MCMC. Additionally, incorporating covariates such as income or education levels into the proposed model could be considered. For example, modelling the parameters of each mixture component (e.g., means, variances) as functions of the covariates. The model in Chapter 4 works well in clustering neighbouring areas that have similar characteristics together in the same cluster smoothly, but it does not always produce fully contiguous clusters, so additional contiguity penalty terms could be considered. Also, as the model in Chapter 4 works for spatial data with single-time points, there is potential to expand this model from univariate to multivariate to fit spatio-temporal data, which has multiple time points. Fitting this extended model to the first step of our proposed approach in Chapter 6 is worth considering.

As the information about disease risk is usually available over a period of time, Chapter 5 presented a two step approach for clustering and estimated disease risk for spatio-temporal data. In this approach, the disease risk is estimated by fitting a spatio-temporal generalised additive mixed model and the clustering data are obtained by using the P-splines' coefficients of the interaction between space and time. A lower-dimensional representation of the underlying data compared to the original and reduce the impact of noise. Clustering the coefficients of the P-splines of the interaction between space and time represents a new application of model-based clustering to epidemiological spatio-temporal data analysis. One of the drawbacks of this approach is some information might be lost during the smoothing transformation to spline coefficients, which will depend on the nature of the data and the goals of the analysis. The proposed model in Chapter 5 assumes no data are missing. In the case of missing data, one possibility is to use imputation. Bayesian approaches are useful when dealing with missing data, as it offers a natural way to consider the uncertainty from missing data when making inference on incomplete data (Ma & Chen, 2018).

The approach in Chapter 6 took into account the potential clustering structure in spatial data over time by adding a cluster factor to the spatio-temporal generalised additive mixed model introduced in Chapter 5. Although generalized additive mixed models (GAMMs) with and without a cluster factor have similar accuracy in disease risk estimation and the difference in the AIC values of both models is quite small, the slight preference for our proposed model based on AIC values suggests that our proposed model, which incorporates clusters, may offer some advantages in capturing underlying patterns and essential information in the datasets. For the models in Chapter 5 and Chapter 6, various alternative types of smoothing could be used depending on the goals of the study.

In this thesis, the simulated data were generated using the same model as described in each respective chapter. In future work, it would be beneficial to conduct more extensive simulations, which would provide further insights into the model's behaviour under various conditions. Also, well-known information criteria such as the DIC and BIC were used to choose the final number of clusters. Future work should consider sensitivity analyses on simulation studies with different number of clusters, to see how these criteria perform.

Bibliography

Adin, A., Lee, D., Goicoa, T. & Ugarte, M. D. (2019), 'A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters', *Statistical Methods in Medical Research* **28**(9), 2595–2613.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on Automatic Control* **19**(6), 716–723.

Alfó, M., Nieddu, L. & Vicari, D. (2009), 'Finite mixture models for mapping spatially dependent disease counts', *Biometrical Journal* **51**(1), 84–97.

Anderson, C., Lee, D. & Dean, N. (2014), 'Identifying clusters in Bayesian disease mapping', *Biostatistics* **15**(3), 457–469.

Anderson, C., Lee, D. & Dean, N. (2016), 'Bayesian cluster detection via adjacency modelling', *Spatial and Spatio-temporal Epidemiology* **16**.

Bayes, T. (1763), 'An essay towards solving a problem in the doctrine of chances', *Phil. Trans. of the Royal Soc. of London* **53**, 370–418.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995), 'Bayesian analysis of space—time variation in disease risk', *Statistics in Medicine* **14**(21-22), 2433–2443.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 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.

Bivand, R. S., Pebesma, E. J. & Gomez-Rubio, V. (2008), *Applied spatial data analysis with R*, Springer, London;New York;.

Blekas, K., Likas, A., Galatsanos, N. P. & Lagaris, I. E. (2005), 'A spatially constrained mixture model for image segmentation', *IEEE transactions on Neural Networks* **16**(2), 494–498.

Böhning, D., Dietz, E. & Schlattmann, P. (2000), 'Space-time mixture modelling of public health data', *Statistics in Medicine* **19**(17-18), 2333–2344.

Casella, G. & Berger, R. (2002), *Statistical Inference*, Duxbury advanced series in statistics and decision sciences, Thomson Learning.

Celeux, G. & Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern Recognition* **28**(5), 781–793.

Charras-Garrido, M., Abrial, D., Goër, J. D., Dachian, S. & Peyrard, N. (2012), 'Classification method for disease risk mapping based on discrete hidden Markov random fields', *Biostatistics* **13**(2), 241–255.

Chavent, M., Kuentz-Simonet, V., Labenne, A. & Saracco, J. (2018), 'Clustgeo: an R package for hierarchical clustering with spatial constraints', *Computational Statistics* **33**(4), 1799–1822.

Congdon, P. (2007), 'Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes', *Computational Statistics & Data Analysis* **51**(6), 3197–3212.

Cressie, N. (1993), *Statistics for spatial data*, Wiley series in probability and mathematical statistics: Applied probability and statistics, J. Wiley.

Davis, F. W. (1993), Introduction to spatial statistics, *in* S. A. Levin, T. M. Powell & J. W. Steele, eds, 'Patch Dynamics', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 16–26.

de Boor, C. (1972), 'On calculating with b-splines', *Journal of Approximation Theory* **6**(1), 50–62.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Dobson, A. J. & Barnett, A. G. (2018), *An introduction to generalized linear models*, CRC press.

Eilers, P. & Marx, B. (1996), 'Flexible smoothing with *b*-splines and penalties', *Statistical Science* **11**(2), 89–102.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association* **97**(458), 611–631.

Fruhwirth-Schnatter, S., Celeux, G. & Robert, C. P. (2019), *Handbook of Mixture Analysis*, CRC Press.

Gaetan, C. & Guyon, X. (2010), *Spatial Statistics and Modeling*, Springer Series in Statistics, Springer Science+Business Media, LLC, New York, NY.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, third edn, CRC Press, Boca Raton, FL.

Gelman, A., Rubin, D. B. et al. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Giordani, P. (2019), *INTRODUCTION TO CLUSTERING WITH R*, SPRINGER, Place of publication not identified.

Grazian, C. (2023), 'Spatio-temporal stick-breaking process', *arXiv preprint arXiv:2303.17177* .

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Green, P. J. & Richardson, S. (2002), 'Hidden Markov models and disease mapping', *Journal of the American Statistical Association* **97**(460), 1055–1070.

Hastie, T. & Tibshirani, R. (1987), 'Generalized additive models: some applications', *Journal of the American Statistical Association* **82**(398), 371–386.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Health, S. & Data, S. C. O. (2020), Seven day trends by neighbourhood, Technical report.

Health, S. & Data, S. C. O. (2021), Daily case trends by health board, Technical report.

Held, L. & Sabanés Bové, D. (2014), 'Applied statistical inference', *Springer, Berlin Heidelberg, doi* **10**(978-3), 16.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, 1st edn, Springer Publishing Company, Incorporated.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of Classification* **2**, 193–218.

Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation problems', *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186**(1007), 453–461.

Knorr-Held, L. (2000), 'Bayesian modelling of inseparable space-time variation in disease risk', *Statistics in Medicine* **19**(17-18), 2555–2567.

Knorr-Held, L. & Rasser, G. (2000), 'Bayesian detection of clusters and discontinuities in disease maps', *Biometrics* **56**(1), 13–21.

Kulldorff, M. (1997), 'A spatial scan statistic', *Communications in Statistics-Theory and methods* **26**(6), 1481–1496.

Lawson, A. B., Choi, J., Cai, B., Hossain, M., Kirby, R. S. & Liu, J. (2012), 'Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data', *Journal of Agricultural, Biological, and Environmental Statistics* **17**, 417–441.

Lee, D. (2011), 'A comparison of conditional autoregressive models used in Bayesian disease mapping', *Spatial and Spatio-temporal Epidemiology* **2**, 79–89.

Lee, D., Rushworth, A. & Napier, G. (2018), 'Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package', *Journal of Statistical Software* **84**(9), 1–39.

Leroux, B. G., Lei, X. & Breslow, N. (2000), Estimation of disease rates in small areas: A new mixed model for spatial dependence, *in* M. E. Halloran & D. Berry, eds,

'Statistical Models in Epidemiology, the Environment, and Clinical Trials', Springer New York, NY, pp. 179–191.

Li, G., Best, N., Hansell, A. L., Ahmed, I. & Richardson, S. (2012), 'BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice', *Biostatistics* **13**(4), 695–710.

Lin, X. & Zhang, D. (1999), 'Inference in generalized additive mixed models by using smoothing splines', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **61**(2), 381–400.

Ma, Z. & Chen, G. (2018), 'Bayesian methods for dealing with missing data problems', *Journal of the Korean Statistical Society* **47**, 297–313.

MacNab, Y. C. & Dean, C. (2001), 'Autoregressive spatial smoothing and temporal spline smoothing for mapping rates', *Biometrics* **57**(3), 949–956.

MacNab, Y. C., Kmetic, A., Gustafson, P. & Sheps, S. (2006), 'An innovative application of Bayesian disease mapping methods to patient safety research: A Canadian adverse medical event study', *Statistics in Medicine* **25**(23), 3960–3980.

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E. & Roser, M. (2020), 'Coronavirus pandemic (covid-19)', *Our World in Data* . https://ourworldindata.org/coronavirus.

McLachlan, G. J. & Peel, D. (2000), *Finite mixture models*, Wiley, New York;Chichester;.

McLachlan, G. & Krishnan, T. (1997), *The EM algorithm and extensions*, Wiley, New York.

Menni, C., Sudre, C. H., Steves, C. J., Ourselin, S. & Spector, T. D. (2020), 'Quanti-

fying additional covid-19 symptoms will save lives', *The Lancet* **395**(10241), e107–e108.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Napier, G., Lee, D., Robertson, C. & Lawson, A. (2019), 'A Bayesian space–time model for clustering areal units based on their disease trends', *Biostatistics* **20**(4), 681–697.

Napier, G., Lee, D., Robertson, C., Lawson, A. & Pollock, K. G. (2016), 'A model to estimate the impact of changes in mmr vaccine uptake on inequalities in measles susceptibility in scotland', *Statistical Methods in Medical Research* **25**(4), 1185–1200.

Nelder, J. A. & Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society Series A: Statistics in Society* **135**(3), 370–384.

Neyman, J. (1937), 'Outline of a theory of statistical estimation based on the classical theory of probability', *Philosophical transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences* **236**(767), 333–380.

NHS-Scotland (2016), 'Health inequalities: What are they? how do we reduce them'.

O'Sullivan, F. (1986), 'A statistical perspective on ill-posed inverse problems', *Statistical Science* pp. 502–518.

O'Sullivan, F. (1988), 'Fast computation of fully automated log-density and log-hazard estimators', *SIAM Journal on Scientific and Statistical Computing* **9**(2), 363–379.

Pearson, K. (1894), 'Contributions to the mathematical theory of evolution', *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110.

R Core Team, R. et al. (2013), 'R: A language and environment for statistical computing'.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association* **66**(336), 846–850.

Richardson, S. & Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **59**(4), 731–792.

Roberts, G. O. & Rosenthal, J. S. (1998), 'Optimal scaling of discrete approximations to langevin diffusions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 255–268.

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.

Rushworth, A., Lee, D. & Mitchell, R. (2014), 'A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London', *Spatial and Spatio-temporal Epidemiology* **10**, 29–38.

Sanjay-Gopal, S. & Hebert, T. J. (1998), 'Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm', *IEEE Transactions on Image Processing* **7**(7), 1014–1028.

Santafé, G., Adin, A., Lee, D. & Ugarte, M. D. (2021), 'Dealing with risk discontinuities to estimate cancer mortality risks when the number of small areas is large', *Statistical Methods in Medical Research* **30**(1), 6–21.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Scottish Government (2020), Coronavirus (covid-19) confirmed in Scotland, Technical report.

Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. (2016), 'mclust 5: clustering, classification and density estimation using Gaussian finite mixture models', *The R Journal* **8**(1), 289–317.

Snow, J. (1855), 'On the mode of communication of cholera, John Churchill'.

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.

Stephens, M. (2000), 'Dealing with label switching in mixture models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 795–809.

Stern, H. S. & Cressie, N. (2000), 'Posterior predictive model checks for disease mapping models', *Statistics in Medicine* **19**(17-18), 2377–2397.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the Detroit region', *Economic Geography* **46**(sup1), 234–240.

Ugarte, M., Goicoa, T. & Militino, A. (2010), 'Spatio-temporal modeling of mortality risks using penalized splines', *Environmetrics: The Official Journal of the International Environmetrics Society* **21**(3-4), 270–289.

Wakefield, J. & Kim, A. (2013), 'A Bayesian model for cluster detection', *Biostatistics* **14**(4), 752–765.

Waller, L. A., Carlin, B. P., Xia, H. & Gelfand, A. E. (1997), 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**(438), 607–617.

Wang, W., Xiao, X., Qian, J., Chen, S., Liao, F., Yin, F., Zhang, T., Li, X. & Ma, Y. (2022), 'Reclaiming independence in spatial-clustering datasets: A series of data-driven spatial weights matrices', *Statistics in Medicine* **41**(15), 2939–2956.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R, Second Edition*, Chapman & Hall/CRC Texts in Statistical Science, CRC Press.

World Health Organization (2020*a*), Novel coronavirus (2019-ncov) situation report-1, Technical report.

World Health Organization (2020*b*), Novel coronavirus (2019-ncov) situation report-71, Technical report.

World Health Organization (2020*c*), Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020, Technical report.

World Health Organization (2020*d*), Who director-general's statement on ihr emergency committee on novel coronavirus (2019-ncov), Technical report.

Wu, F.-Y. (1982), 'The Potts model', *Reviews of Modern Physics* **54**(1), 235.

Yin, X., Napier, G., Anderson, C. & Lee, D. (2022), 'Spatio-temporal disease risk estimation using clustering-based adjacency modelling', *Statistical Methods in Medical Research* **31**(6), 1184–1203.

Zacks, S. (1981), *Parametric statistical inference: basic theory and modern approaches*, Vol. 4;4.;, Pergamon Press, New York;Oxford;.