Risk estimation and discontinuity identification in Bayesian disease mapping

Xueqing Yin

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

August 2022

Abstract

Disease mapping is the field of epidemiology that estimates the spatial or spatio-temporal pattern in disease risk. Approaches in this field are generally based on data collected on a set of non-overlapping areal units that comprise the study region, and typically utilise counts of the numbers of disease cases within each areal unit. Conditional autoregressive (CAR) models are commonly used to capture the spatial autocorrelation present in areal unit disease count data. The spatial correlation structure that is induced by these models is typically determined by a neighbourhood matrix based on geographical adjacency, which enforces spatial correlation between geographically neighbouring areas and assumes a spatially smooth risk surface. However this may not be realistic in practice, because some pairs of neighbouring areas are likely to exhibit vastly different disease risks. Therefore the aim of this thesis is to develop methodology that allows for discontinuities in the spatial risk pattern when estimating disease risk. The first two models proposed are in a purely spatial setting and account for discontinuities by identifying spatial clusters of areas that have higher or lower risks than their geographical neighbours, while the third proposed model extends this to the spatio-temporal domain to identify clusters/discontinuities and estimate the spatial pattern of disease risk over time. The final piece of work of this thesis allows for discontinuities by using a boundary analysis approach. This approach identifies the boundaries in the spatial risk surface that separate pairs of geographically adjacent areas that exhibit large differences between their risks. Each model is applied to hospital admissions data for respiratory disease from the Greater Glasgow and Clyde Health Board region. Overall, it has been found that the respiratory disease risk surface in Greater Glasgow is not globally spatially smooth. There are numerous pairs of neighbouring areas where a discontinuity in disease risk appears to exist. In addition, the respiratory disease risk in Glasgow appears to increase over time and people living in more deprived areas are at higher risk of respiratory hospital admissions than those living in more affluent areas.

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Prof Duncan Lee, Dr Craig Anderson and Dr Gary Napier for their invaluable guidance, constant support and encouragement throughout each stage of my PhD study. Their extensive knowledge and plentiful experience have inspired me in all the time of my research and writing of this thesis. Without them, this work would not have been possible. I also gratefully acknowledge the PhD scholarship received from University of Glasgow.

I deeply thank my family, in particular, my parents for their unconditional trust and support. Without their love and understanding in the past few years, it would be impossible for me to complete this journey. Thank you to Yingjie, for his support, never-ending encouragement when times are tough, and for always keeping me cheerful. I would also like to thank my friends, with whom I have shared moments of deep anxiety but also of big excitement. Their presence makes my study and life in Glasgow a wonderful time.

Declaration

I hereby declare that this thesis has been written by me, and has not been submitted previously as part of any application for a degree. I confirm that the work presented in this thesis is my own, except where otherwise explicitly stated by reference or acknowledgment.

The work presented in Chapter 4 was presented at the 13_{th} CFE-CMStatistics virtual conference in 2020. The work presented in Chapter 5 has been published in the Statistical Methods in Medical Research journal with the title “Spatio-temporal disease risk estimation using clustering-based adjacency modelling”, and is jointly authored by Dr Gary Napier, Dr Craig Anderson and Prof Duncan Lee. I delivered an invited talk (virtually) on this work at the 14_{th} CFE-CMStatistics conference in London, UK in 2021, and also at the 44_{th} Research Students’ Conference in Probability and Statistics in 2021.

Chapter 1

Introduction

Disease mapping is the field of statistical epidemiology that studies the spatial or spatio-temporal distribution of population-level disease risk (Lawson et al., 2016). Quantifying the spatial variation in disease risk is of great importance for improving public health, as for example it allows researchers to identify possible underlying factors explaining the variation, as well as allowing health authorities to design and evaluate the effect of public health policies such as the allocation of health care resources. The differences in disease risk across social groups and between different population groups are known as health inequalities (NHS Health Scotland, 2016), which are often related to risk factors such as population behaviours (e.g. smoking, alcohol consumption, exercise) and environmental exposures (e.g. air pollution, water quality) (Lawson and Lee, 2017). Poverty or socio-economic deprivation is one of the key reasons for these differences, with deprived areas being more likely to exhibit higher disease risks than more affluent areas (McCartney, 2012, NHS Health Scotland, 2016). One of the earliest disease mapping studies can be traced back to the work of Seaman (1798), who produced a dot disease map of the locations of residences affected by yellow fever in 1795 in New Slip, New York as shown in Figure 1.1. He mapped the infected cases as well as the local waste sites, and concluded that yellow fever originated in these waste areas. In 1854, London was experiencing a deadly cholera epidemic. At the time, people believed that cholera was caused by bad air. However, Snow (1855) mapped the cholera cases in Soho, London using small stacked bars (see Figure 1.2), and ultimately determined the source of the outbreak was an infected public water pump. John Snow is widely considered to be the father of modern epidemiology for determining how cholera was transmitted and the statistical mapping methods he initiated.

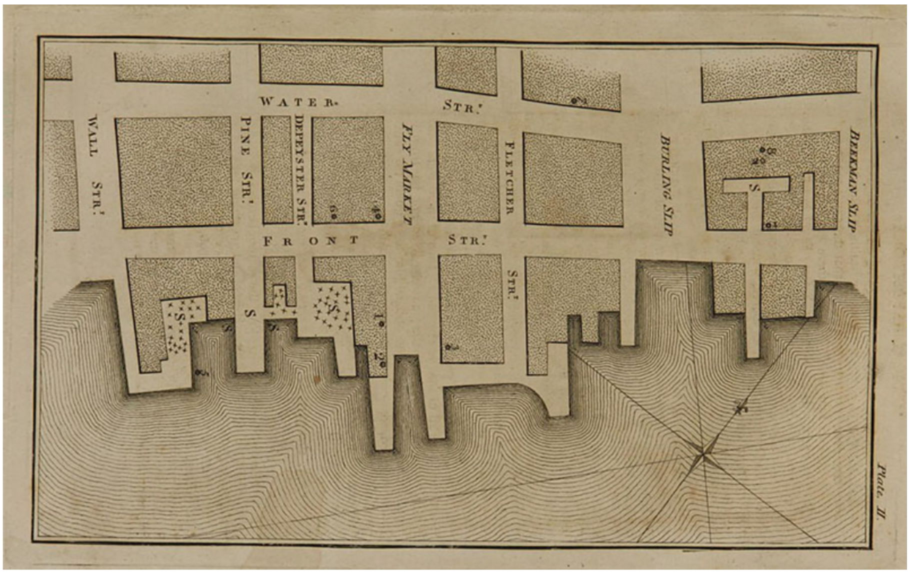


Figure 1.1: Seaman's map of the sources of the 1795 yellow fever outbreak in New York City, with infected cases marked with dots and waste sites marked with S's and crosses. Source: U.S. National Library of Medicine.

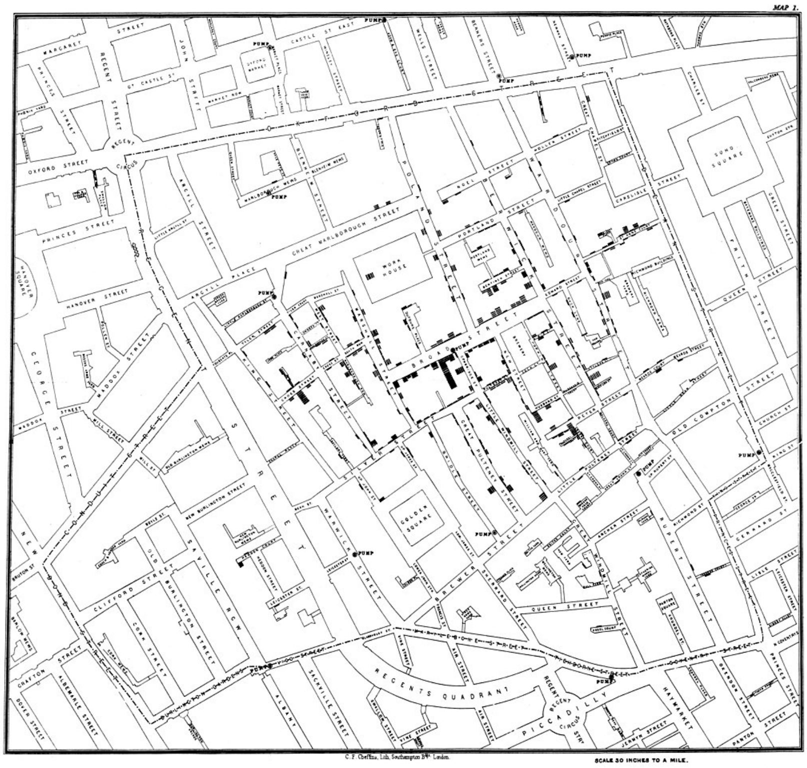


Figure 1.2: The original map by Snow (1855) showing the spread of cholera in Soho, London in 1855, with infected cases marked using small bars.

Modern disease mapping studies usually illustrate the spatial patterns of differences in ill health via disease maps, where areas are shaded on a scale in different colours relating to disease risk. Most disease maps utilise data collected on non-overlapping areal units that comprise the study region, such as electoral wards, census tracts or health board areas. This is because individual level data can not be made open to the public due to confidentiality reasons, therefore only aggregated data such as counts of the numbers of disease cases or mortality from the population living in each area are available. Disease counts will depend on the overall size and demographic structure (e.g. age and sex) of the populations living within each area, thus an age and sex adjusted measure of disease risk is required to put the disease risks for different areas on the same scale for proper comparison. This is done by computing the expected disease counts in each area via indirect standardisation based on its population demographics and national age-sex specific disease rates. The simplest measure of disease risk is the standardised incidence ratio (SIR), which is the ratio of the observed counts to the expected counts of disease cases for each areal unit. Values of SIR less than 1 indicate lower levels of disease risk compared to expected, while values greater than 1 correspond to higher levels of disease risk than expected. For example, an SIR of 1.2 corresponds to a risk 20% higher than expected for that area. However, the SIR is an unstable and sometimes uninformative estimate of disease risk, especially when the population of the study is small or the disease in question is rare, in which case some areas may have small values of the expected number of disease cases. Additionally, the naive mapping of SIR ignores the spatial autocorrelation which could be present in the data, and the SIR also ignores the estimation of covariate effects. These issues thus have led researchers to instead estimate disease risk patterns using model-based approaches.

Two inferential goals are relevant to modern disease mapping studies (Gelfand et al., 2010): (i) computing precise estimates of disease risk in small areas; (ii) identifying high/low-risk areas. Bayesian hierarchical models have been commonly adopted during the last years to deal with these goals, by incorporating a set of spatially varying random effects that account for the spatial autocorrelation in the data. These random effects are typically modelled via a conditional autoregressive (CAR) model (Besag et al., 1991, Leroux et al., 2000), which assumes spatial autocorrelation between neighbouring areas based on the idea that *"Everything is related to everything else, but near things are more related than distant things"* (Tobler, 1970). This correlation is represented by a neighbourhood matrix $\boldsymbol{W}$, which summarises the spatial closeness between each pair of areal units. Typically $\boldsymbol{W}$ contains binary elements $\{w_{ij}\}$ whose values are determined based on a border sharing specification, where $w_{ij} = 1$ if areal units (i, j) share a common geographical border and $w_{ij} = 0$ otherwise. Such CAR models smooth the SIRs in each area by borrowing information from neighbouring areas to remove random noise, and as a result, the estimated spatial risk pattern is globally spatially smooth. However, these models may not be appropriate if the goal is instead to identify clusters of high-risk areas or boundaries of rapid changes in the risk surface, because they assume a constant level of spatial autocorrelation across the entire study region and so force neighbouring areas to have similar risks, which may prevent the identification of high-risk areas as discontinuities in the smooth risk become blurred. This can also lead to poorer risk estimation because in practice, the level of spatial autocorrelation can vary across the study region. Some pairs of neighbouring areas may display strong spatial smoothness and have similar risks, while other pairs exhibit weak or no spatial autocorrelation and have vastly different disease risks. Therefore enforcing a constant level of spatial autocorrelation across the study region may result in incorrect spatial smoothing of disease risk between some neighbouring areas and hence reduce the accuracy of risk estimation. Numerous approaches have been developed to allow for spatial discontinuities in the disease risk surface, by identifying either clusters of areas that exhibit elevated or reduced risks compared to their neighbours (Knorr-Held and Raßer, 2000, Anderson et al., 2014, Adin et al., 2019, Santafé et al., 2021) or boundaries that separate two geographically adjacent areas with largely different risks (Lu and Carlin, 2005, Lu et al., 2007, Lee and Mitchell, 2013, Lee et al., 2014, Rushworth et al., 2017, Lee et al., 2021). The identification of spatial discontinuities in the data is crucial for social epidemiologists and health policy makers. On the one hand, knowledge of the geographical extent of high-risk clusters enables disease prevention and health funding to be targeted appropriately. On the other hand, the locations of boundaries are likely to represent the demarcation between different communities or neighbourhoods, and reflect changes in the underlying biological, physical or social processes (Lee et al., 2021, Jacquez et al., 2000).

Estimating risks via spatial smoothing and detecting risk discontinuities are two contradictory goals in disease mapping. Therefore, the main aim of this thesis is to develop spatial and spatio-temporal methodology that can achieve a trade-off between smoothing and identifying discontinuities in the spatial pattern of disease risk. In this thesis, models are built based on aggregated disease count data at the areal unit level, and the study region considered here is the Greater Glasgow and Clyde Health Board in West Central Scotland, which provides health care to a population of around 1.2 million people (`https://www.nhsggc.org.uk/`). This region is the largest health board in both Scotland and the UK in terms of population size, which consists of Glasgow, the largest city in Scotland, and the surrounding areas including East Dunbartonshire, East Renfrewshire, Inverclyde, Renfrewshire and West Dunbartonshire. The small areal units for which disease data are available are known as Intermediate Zones (IZs) (`https://statistics.gov.scot/home`). The Greater

Glasgow and Clyde Health Board contains a total of 257 IZs, whose populations range between 2,500 and 6,000 in a single IZ with an average population of approximately 4,000 residents. The geographical sizes of these 257 IZs are different and depend on the population density of each IZ. Figure 1.3 presents the map of the IZs in the Greater Glasgow study region, and shows that the densely populated areas in the center of the health board have much smaller geographical size than the more rural areas with sparse populations.

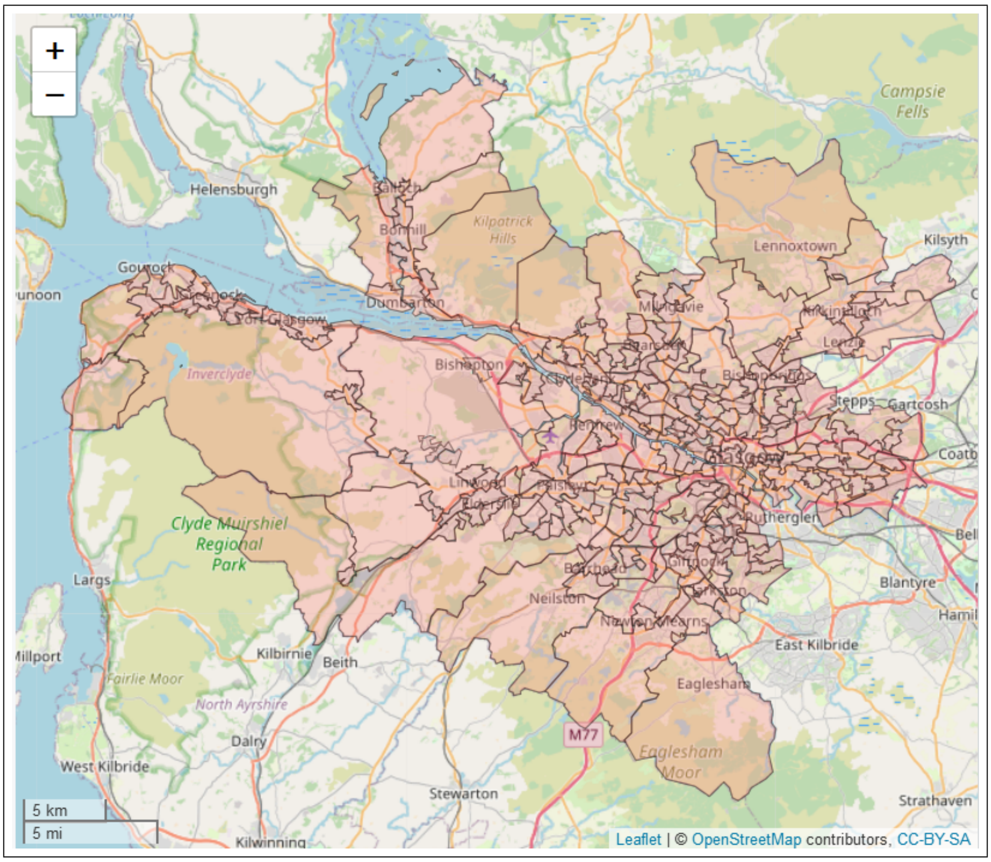


Figure 1.3: A map of the 257 Intermediate Zones (IZs) in the Greater Glasgow and Clyde Health Board.

The Greater Glasgow and Clyde Health Board is selected due to a few reasons. Firstly, Glasgow exhibits some of the poorest health profiles compared to the rest of the United Kingdom and western Europe, which is known as the *"Glasgow effect"* (Walsh et al., 2010). This can be seen in Figure 1.4, which shows the male life expectancy at birth in Glasgow compared to other major cities in the UK from 1991-2020. The life expectancy estimates are calculated using abridged period life tables constructed by the established methodology developed by Chiang et al. (1979). Abridged life tables aggregate deaths and population data into age intervals, e.g. 0-1, 1-4, 5-9, . . . , 80-84, 85 over. Age-specific mortality rates are used within the life table to calculate the probability of dying at each age interval. These probabilities are then applied to a hypothetical population cohort of newborn babies. The life expectancy at birth in an area can thus be defined as an estimate of the average number of years a new-born baby would survive if he or she experienced the age-specific mortality rates of the given area and time period throughout his or her life, which is computed as the total number of person-years lived beyond exact age 0 divided by the number of newborns. One disadvantage of such life table estimates is that they do not account for the migratory segments of individuals during their lifetime (Chiang et al., 1979). Figure 1.4 shows that although the male life expectancy in Glasgow has been improving overall over the 29 year period, it remains the lowest of all major cities in the UK and the gaps between Glasgow and the other cities are still widening. The figure also illustrates that improvements in male life expectancy appear to level off in recent years. A sharp reduction in life expectancy for men across UK cities was observed in 2018-2020, which is likely to result from the increased mortality caused by the Covid-19 pandemic occurring in early 2020. Figure 1.5 displays the estimated male life expectancy across Glasgow neighbourhoods between 2015 and 2019, where large health inequalities can be clearly seen. The life expectancy of males living in Glasgow ranged from 65.4 years in Govan (a deprived area) to 83 years in Pollokshields West (an affluent area), which is a gap of around 18 years.

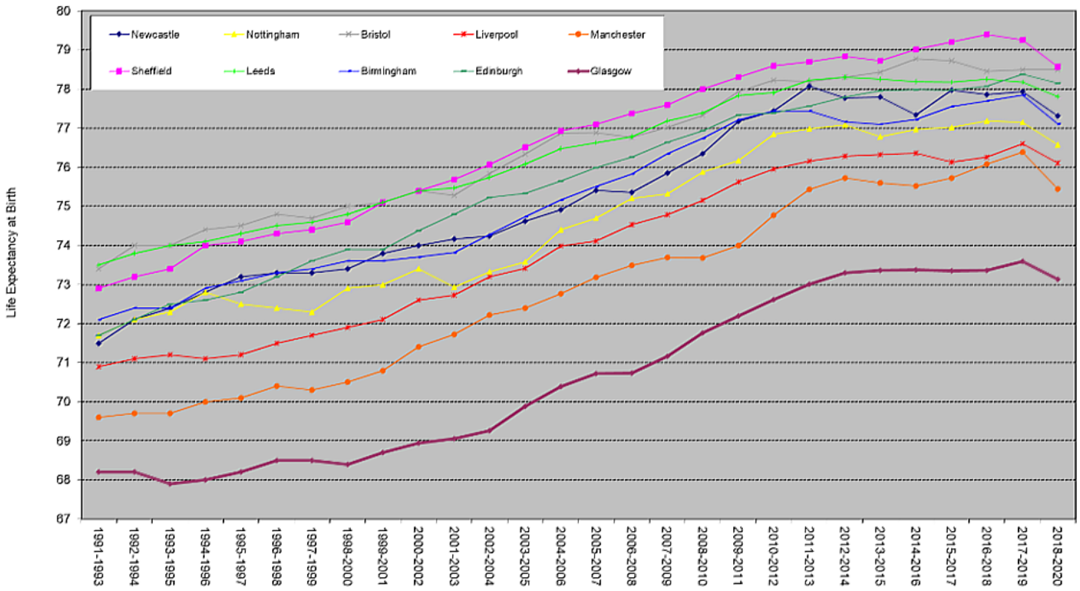


Figure 1.4: Male life expectancy for Glasgow compared to other selected cities in the UK from 1991-93 to 2018-20. Plot source: The Glasgow Centre for Population Health (2021). Data source: ONS, National Records of Scotland.

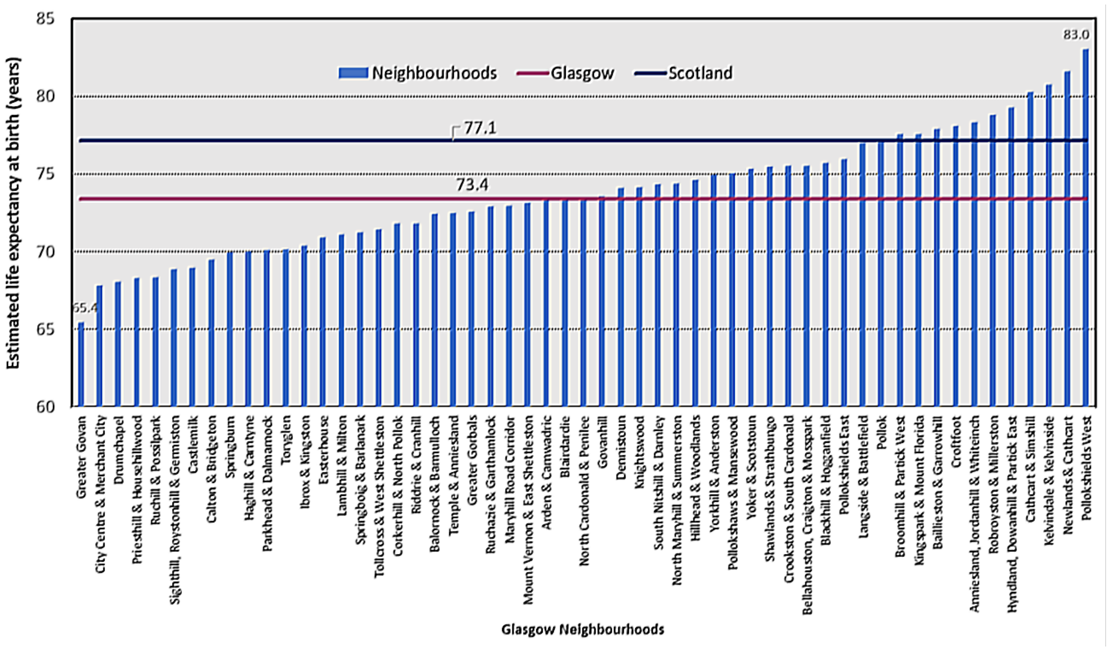


Figure 1.5: Estimated male life expectancy across Glasgow neighbourhoods between 2015 and 2019. The estimates are based on Chiang II methodology (Chiang et al., 1979). Plot source: The Glasgow Centre for Population Health (2021). Data source: National Records of Scotland.

The methodology proposed in this thesis is applied to hospital admissions data for respiratory disease (defined using the International Classification of Disease tenth revision by codes J00-J99 and R09.1) in the Greater Glasgow and Clyde Health Board region. Respiratory disease is selected due to it being a significant cause of morbidity and mortality in the UK, which accounts for 6.5% of hospital admissions and 24% of all deaths (Chung et al., 2002, Lozano et al., 2012). Chronic obstructive pulmonary disease (COPD), lower and upper respiratory tract infections, occupational lung diseases, reactive airway diseases and lung cancer are the leading causes of hospital admissions due to respiratory disease in the UK, which lead to serious public health problems and financial burden (Salciccioli et al., 2018). Factors such as smoking, air pollution from traffic and industrial sources, low socioeconomic status are important contributors to most respiratory conditions (Forum of International Respiratory Societies, 2017, Troeger et al., 2018).

I develop spatial and spatio-temporal modelling approaches for estimating disease risk and identifying discontinuities in the risk surface in Chapters 3, 4, 5 and 6, the first three of which identify risk discontinuities by partitioning the entire study region into disjoint clusters of areas exhibiting higher or lower risks than their geographical neighbours, while the last chapter accounts for discontinuities by identifying the locations of boundaries where geographically neighbouring areas have very different disease risks. In Chapter 3, a clustering method (k-means clustering) is applied to data to produce a set of candidate cluster structures, and then a separate Bayesian hierarchical model is fitted for each candidate structure. The most appropriate cluster structure is chosen by model comparison techniques, including the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and the effective number of independent parameters. In order to provide more flexibility in cluster identification as well as quantifying the uncertainty in the cluster structure, in Chapter 4 I construct a much bigger set of candidate cluster structures using multiple clustering methods, and fit a single spatial model in a Bayesian setting, with the optimal cluster structure estimated as a parameter within that model. This spatial modelling approach is then extended to the spatio-temporal modelling of data in Chapter 5 to estimate disease risk and clusters over time, where the spatial clusters either remain fixed or evolve dynamically over time. These modelling approaches have a common feature of allowing the disease risk to be correlated for pairs of geographically neighbouring areal units within the same cluster, but conditionally independent for neighbouring areas in different clusters.

In Chapter 6 I use a boundary analysis approach to identify discontinuities in the spatial risk pattern. Rather than seeking for spatial clusters of areas identified as being similar or different in disease risk as in the previous chapters, here the differences in disease risk between each pair of geographically adjacent areal units are of interest. These differences are known as “boundaries” (large or small) in the risk surface, which can be open and do not necessarily completely enclose an area or group of areal units. I adopt the graph-based optimisation algorithm proposed by Lee et al. (2021) to estimate a number of candidate neighbourhood matrices, which represent a range of potential boundary structures for the data. The algorithm treats each element w_{ij} in the neighbourhood matrix $\boldsymbol{W}$ as a binary random quantity if areal units (i, j) share a common geographical border. If w_{ij} is eventually estimated as 0 then a boundary is said to exist between areas i and j in the risk surface, while an estimate of $w_{ij} = 1$ suggests no risk boundary between the two areas. Then a spatio-temporal model is fitted to the data, which jointly estimates disease risk over time and identifies the locations of boundaries by estimating the neighbourhood matrix for the data. This model does not allow for the smoothing of disease risk between pairs of neighbouring areas that have a boundary.

The remainder of this thesis is organised as follows. Chapter 2 provides an overview of the existing statistical methodology which is used throughout this thesis, and the related literature which particularly focuses on spatial and spatio-temporal modelling, as well as spatial clustering and boundary analysis in disease mapping. In Chapters 3 and 4, two spatial modelling approaches are developed for estimating disease risk and identifying the spatial cluster structure. The approach presented in Chapter 4 is extended to the spatial-temporal domain in Chapter 5, which either forces the clusters/discontinuities to be the same for all time periods or allows them to evolve dynamically over time. Chapter 6 discusses an approach that identifies the locations of boundaries corresponding to large differences in disease risk between neighbouring areas. Finally, Chapter 7 summarises the work contained in the thesis and discusses possible avenues for future research. The spatial methodology presented in Chapters 3 and 4 is applied to hospital admissions data for respiratory disease for 2016 across the Greater Glasgow and Clyde Health Board, while the spatio-temporal methodology presented in Chapters 5 and 6 is applied to respiratory disease data for the years 2011-2017 for the same health board region. Inference in all the proposed models is carried out in a Bayesian setting. Inference for the models in Chapters 3, 4 and 5 is based on Markov chain Monte Carlo (MCMC) simulation, while model inference in Chapter 6 is based on a Metropolis-coupled Markov chain Monte Carlo algorithm due to the multimodality issue. All the models are developed and written in R (R Core Team, 2013), while some complex parts are developed in the more efficient C++ language via the Rccp package (Eddelbuettel et al., 2011).

Chapter 2

Literature review

This chapter provides an overview of the statistical theory and methodology which are used throughout this thesis as well as the related literature. Section 2.1 introduces Bayesian statistics, which is the statistical framework utilised in this thesis. Section 2.2 introduces generalised linear models (GLMs), and particularly focuses on the Poisson GLMs for count data. Section 2.3 introduces some commonly used model selection techniques when comparing multiple statistical models. Section 2.4 gives a brief outline of spatial statistics, focusing on the field of spatial modelling of areal unit data in a disease context, which is known as disease mapping. This section forms the basis of the spatial methodology developed in Chapters 3 and 4. Some of the existing spatio-temporal disease mapping literature is explored in Section 2.5, which forms the basis of Chapters 5 and 6. Section 2.6 explores existing approaches which allow for discontinuities in the spatial risk pattern in a disease mapping context, which is also one of the focuses of this thesis. Finally, Section 2.7 outlines a range of clustering approaches which form part of the methodology developed in Chapters 3, 4 and 5, and Section 2.8 introduces the receiver-operating characteristic (ROC) curves and the area under the ROC curve which are used to quantify the accuracy of the boundary identification in Chapter 6.

2.1 Bayesian statistics

2.1.1 Introduction

Bayesian statistics is a branch of statistics that helps people update their prior beliefs about random events to produce new posterior beliefs after being given additional information such as new data or related evidence. Bayes' theorem (Bayes, 1763) was developed by Thomas

Bayes in the 18th century and is stated as follows for random events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A)$ and $P(B)$ are the probabilities of events A and B occurring, and $P(A|B)$ is the conditional probability of event A occurring given B has occurred. There are two schools of statistical inference, namely frequentist statistics and Bayesian statistics. In frequentist statistics, the observed data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are treated as random samples whereas the unknown underlying parameters of interest $\boldsymbol{\theta}$ are assumed fixed. Under this framework, parameters $\boldsymbol{\theta}$ are typically estimated from the data $\boldsymbol{Y}$ by maximizing the likelihood function denoted by $L(\boldsymbol{\theta}|\boldsymbol{Y}) = \prod_{i=1}^{n} f(Y_i|\boldsymbol{\theta})$, where $f(Y_i|\boldsymbol{\theta})$ is the probability distribution function of Y_i. The uncertainty of a point estimate (e.g. the maximum likelihood estimator) is measured by a $a\%$ confidence interval. For example, a 95% confidence interval means that if the data are repeatedly sampled for 100 times with a confidence interval constructed each time, then 95% of the resulting 100 confidence intervals will be expected to contain the true value of the parameter. Unlike frequentist statistics, in Bayesian statistics the observed data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are assumed to remain fixed whereas the model parameters $\boldsymbol{\theta}$ are treated as random variables which are estimated in terms of probability statements. Each parameter can be assigned a probability distribution in advance, which is known as a prior distribution $f(\boldsymbol{\theta})$. A prior distribution expresses our prior belief about a model parameter, which can then be updated based on the observed data $\boldsymbol{Y}$ to determine a posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y})$ as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{Y})}. \tag{2.1}$$

Here $f(\boldsymbol{\theta}|\boldsymbol{Y})$ is the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{Y}$, $f(\boldsymbol{Y}|\boldsymbol{\theta})$ is the data likelihood and $f(\boldsymbol{Y})$ is the marginal distribution of the data. If $\boldsymbol{\theta}$ are discrete variables then $f(\boldsymbol{Y}) = \sum_{\boldsymbol{\theta}} f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$, otherwise $f(\boldsymbol{Y}) = \int f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$ for continuous $\boldsymbol{\theta}$. However, since $f(\boldsymbol{Y})$ is usually difficult to estimate and it does not depend on $\boldsymbol{\theta}$, the formula (2.1) can be rewritten up to a constant of proportionality as

$$f(\boldsymbol{\theta}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta}).$$

2.1.2 Prior distributions

A prior distribution $f(\boldsymbol{\theta})$ reflects all of the information we have about the unknown model parameters $\boldsymbol{\theta}$ before observing the data $\boldsymbol{Y}$. This prior distribution is then combined with the data likelihood $f(\boldsymbol{Y}|\boldsymbol{\theta})$ to determine the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y})$. Therefore it is crucial to choose an appropriate prior in order to make sensible inference for parameters. There are various types of prior distribution depending on the available information and the type of data and model. The prior distribution could be specified by using information from previous studies with similar nature or expert knowledge in the field. Such a prior is called an informative prior and has influence on the posterior distribution. However, if little or no information is known about a parameter in advance of the observed data, then a weakly informative prior, which has little influence on the posterior distribution, could be assigned. In this case, the posterior distribution is dominated by the likelihood function and driven by the data rather than the prior. For example, a weakly informative prior for real valued parameters could be a Gaussian distribution with a very large variance, such as $\theta_j \sim$ N(0, 100000), and a weakly informative prior for a parameter on the unit interval could be a uniform prior distribution $\theta_j \sim$ Uniform(0, 1). Another form of weakly informative prior is the Jeffreys prior (Jeffreys, 1946), under which any reparameterisation of the prior is also constant. A Jeffreys prior is formed as

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

where $||\cdot||$ denotes the determinant of a matrix and $I(\boldsymbol{\theta})$ is the Fisher Information defined as

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

The prior that has the same distributional form as the posterior distribution is called a conjugate prior to the likelihood. For example, suppose we have a single observation $Y \sim$ Binomial(n, θ), then the likelihood is

$$L(\theta|Y) = f(Y|\theta) = \binom{n}{Y}\theta^Y(1-\theta)^{n-Y} \propto \theta^Y(1-\theta)^{n-Y}.$$

If a beta distribution is chosen as the prior for θ, i.e. $\theta \sim$ Beta(α, β), then we have

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

The posterior distribution for θ is given as

$$\begin{aligned} f(\theta|Y) &\propto f(Y|\theta)f(\theta) \\ &\propto \theta^{Y}(1-\theta)^{n-Y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}. \end{aligned}$$

Thus, the posterior distribution follows the same parametric form as the prior distribution, that is a beta distribution $f(\theta|Y) \sim \text{Beta}(Y+\alpha, n-Y+\beta)$. Therefore a beta prior distribution is the conjugate prior for the binomial likelihood. Another example of a conjugate prior for the Poisson mean parameter would be a Gamma prior. Conjugate priors are mathematically convenient and thus make inference easier because they allow the posterior to come from a known standard distributional family. A prior is proper if it can integrate to a finite number over its range space, otherwise it is an improper prior. For instance, a uniform prior distribution over all real numbers, $\text{Uniform}(-\infty,\infty)$, is an improper prior. Although an improper prior can lead to a proper posterior distribution, care should be taken because it can also result in an improper posterior which makes inference impossible.

In this thesis, the modelling approaches are developed in a Bayesian setting, and both conjugate and weakly informative priors will be used in the developed models.

2.1.3 Bayesian inference

In order to estimate the parameters in a Bayesian model, Bayesian inference is typically made based on the posterior distributions which are obtained by updating the priors with the data likelihood. Some posterior distributions are straightforward to calculate, for example coming from a standard probability distribution family, then posterior samples can be directly drawn from the distribution of interest. However, in many cases, the posterior distribution is intractable and not easy to derive, so instead more complex and advanced methods are required, which generally draw samples from an approximation of the posterior distribution. Markov chain Monte Carlo (MCMC) simulation is the most commonly used approach when $\boldsymbol{\theta}$ can not be directly sampled from $f(\boldsymbol{\theta}|\boldsymbol{Y})$. MCMC simulation generates a sequence of samples that are approximately drawn from the posterior distribution, by constructing a Markov chain whose converging stationary distribution (or equilibrium distribution) is equal to the target posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y})$ after a number of iterations. In this thesis I make use of two main algorithms to achieve MCMC simulation, which are Gibbs sampling (Geman and Geman, 1984) and the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970).

2.1.3.1 Gibbs sampling

Consider a parameter vector $\boldsymbol{\theta}$ which has been partitioned into d components as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$. The idea in Gibbs sampling is to cycle through each component $\boldsymbol{\theta}_j$ to sample from its full conditional distribution given all the other components fixed to their current values. The Gibbs sampling algorithm for directly drawing S samples from the posterior distribution is as follows.

1. Set initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_d^{(0)})$ to start the Markov chain.

2. At each iteration $i = 1, 2, \ldots, S$ and for each component $j = 1, \ldots, d$, simulate $\boldsymbol{\theta}_j^{(i)}$ in turn from the conditional distribution given all the other components, $f(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j}^{(i)}, \boldsymbol{Y})$, where $\boldsymbol{\theta}_{-j}^{(i)}$ represents the current values of all the components of $\boldsymbol{\theta}$ apart from $\boldsymbol{\theta}_j$, that is

$$\boldsymbol{\theta}_{-j}^{(i)} = (\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{j-1}^{(i)}, \boldsymbol{\theta}_{j+1}^{(i-1)}, \ldots, \boldsymbol{\theta}_d^{(i-1)}).$$

Gibbs sampling works when $f(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j}^{(i)}, \boldsymbol{Y})$ is from a known family of distributions and easy to sample from. For example, if we have a Gaussian conjugate prior for the Gaussian likelihood, then the posterior distribution is also Gaussian, therefore we can directly sample at each step of the Gibbs sampling algorithm. However, when the conditional distribution does not appear to be of any known form and is too intractable to sample from, a more complex method should be considered, such as the Metropolis-Hastings algorithm which involves an acceptance or rejection step.

2.1.3.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is typically used to sample from an intractable target posterior distribution. The Metropolis-Hastings algorithm for drawing S samples proceeds as follows.

1. Set starting values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_d^{(0)})$ to start the Markov chain .

2. At each iteration $i = 1, 2, \ldots, S$ and for each component $j = 1, \ldots, d$, simulate $\boldsymbol{\theta}_j^{(i)}$ using the following steps.

   (a) Simulate a set of proposed parameter values $\boldsymbol{\theta}_j^*$ from a proposal distribution (or jumping distribution), $g(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(i-1)})$.

(b) Calculate the acceptance probability

$$p = \min\left\{ \frac{f(\boldsymbol{\theta}_j^*|\boldsymbol{Y})/g(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_j^{(i-1)})}{f(\boldsymbol{\theta}_j^{(i-1)}|\boldsymbol{Y})/g(\boldsymbol{\theta}_j^{(i-1)}|\boldsymbol{\theta}_j^*)}, 1 \right\}.$$

If the proposal distribution is symmetric, i.e. $g(\boldsymbol{\theta}_j^{(i)}|\boldsymbol{\theta}_j^{(i-1)}) = g(\boldsymbol{\theta}_j^{(i-1)}|\boldsymbol{\theta}_j^{(i)})$ for all $\boldsymbol{\theta}_j^{(i)}$, $\boldsymbol{\theta}_j^{(i-1)}$ and i, this probability can be reduced to the Metropolis algorithm:

$$p = \min\left\{ \frac{f(\boldsymbol{\theta}_j^*|\boldsymbol{Y})}{f(\boldsymbol{\theta}_j^{(i-1)}|\boldsymbol{Y})}, 1 \right\}.$$

(c) Generate a uniform random sample $U \sim \text{Uniform}(0,1)$; if $U \leq p$, accept the proposed $\boldsymbol{\theta}_j^*$ and set $\boldsymbol{\theta}_j^{(i)} = \boldsymbol{\theta}_j^*$; if $U > p$, reject the proposed $\boldsymbol{\theta}_j^*$ and set $\boldsymbol{\theta}_j^{(i)} = \boldsymbol{\theta}_j^{(i-1)}$.

A common choice of the proposal distribution is a Gaussian distribution with mean being the current value $\boldsymbol{\theta}_j^{(i-1)}$ and variance $\boldsymbol{\Sigma}$ because it is symmetric. A proposal distribution with a small variance will make the proposed value close to the current value and so the proposal is more likely to be accepted, resulting in a high acceptance rate. In contrast, if the proposed value is very different from the current value (the proposal distribution has a large variance), it is less likely to be accepted, resulting in a low acceptance rate. A very high acceptance rate indicates that the Markov chain does not explore the full parameter space, whereas a too low acceptance rate means that the Markov chain is stuck on one value for a long period of time. Therefore, we need to choose an appropriate proposal distribution which gives a sensible acceptance rate. Gelman et al. (1996) suggested that the acceptance rate should be around 44% for univariate parameters and around 23% for high-dimensional parameters as this can optimize the efficiency of the algorithm. In this thesis, the acceptance rate will be maintained between 40% and 60% for the parameter of low dimension and between 20% and 40% for parameters of high dimension by tuning the proposal variance every 100 iterations. Note that model parameters can be updated individually in the MCMC simulation, but this would be computationally slow especially when the Bayesian model is complex and has a large number of parameters. One approach to tackle this is to simulate several parameters with similar characteristics at the same time as one block. However, when the block contains a large number of parameters, the sampler may miss some parts of the parameter space and the acceptance rates generally decrease.

The choice of starting values $\boldsymbol{\theta}^{(0)}$ in the Markov chain is likely to affect the performance of the simulation (Brooks, 1998). For example, starting values close to the true values will take a smaller number of iterations to reach the true values than those that are not. Several methods have been proposed for selecting initial values. One method is to run multiple Markov chains with different starting values. Alternatively, estimates from simpler models or methods, e.g. maximum likelihood estimates, can be used as the starting values (Kass et al., 1998).

2.1.3.3 Model convergence

Inference using MCMC simulation is only valid when the Markov chain has converged to the target posterior distribution. The chain will take some time to reach convergence, therefore in order to diminish the influence of posterior samples at early iterations, a general choice is to discard the samples prior to convergence, which is called a burn-in period. The remaining samples, which are assumed to have converged, allow us to calculate the quantities of interest for each of the model parameters.

A number of approaches have been proposed to diagnose the model convergence. One common method is to examine the trace plot of the MCMC samples for each parameter. A trace plot displays the values of a parameter for each iteration in a Markov chain. Figure 2.1 is an example of a trace plot for 2,000 MCMC samples for a parameter. These successive samples are connected by a line to view the path traversed by the chain. Convergence is assumed to have achieved when the trace plot shows no trend and looks weakly stationary. Another graphic tool is to check the trace plot of the running mean for each parameter in the chain, where the running mean is the mean of all sampled values up to a specified number of iterations (Smith, 2005). If the chain has converged then the running mean should stabilize at the posterior mean.

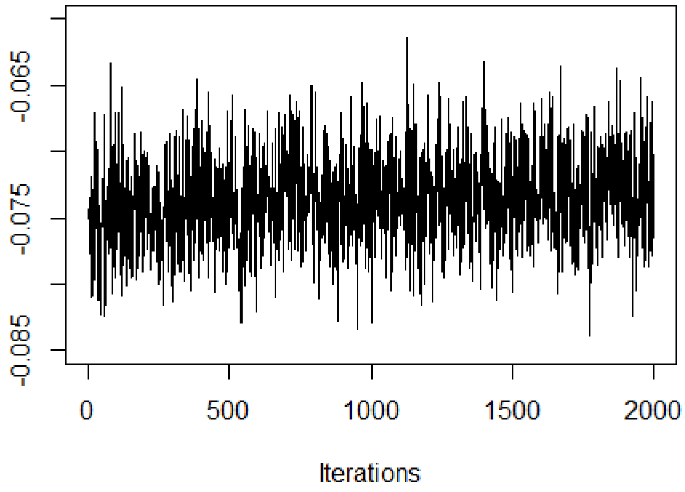


Figure 2.1: An example of a trace plot.

Geweke (1992) proposed to check the model convergence using a Z statistic. The basic idea is to divide a Markov chain into two "windows" containing the first and the last part of the chain, e.g. the first 10% and the last 50%. Ideally, if the chain has reached convergence the means of the two windows should be nearly the same. A Z statistic is calculated as the difference between the two means from these two windows divided by the asymptotic standard error of the difference. As the number of iteration increases, the distribution of the Z statistic approaches the standard normal distribution $\mathrm{N}(0,1)$. Thus a Z statistic with values between [-1.96, 1.96] suggests convergence.

Alternatively, Gelman et al. (1992) proposed to monitor the convergence of multiple and parallel chains based on the within-chain and between-chain variances. Specifically, a potential scale reduction factor (PSRF), which estimates the potential decrease in the between-chains variability with respect to the within-chain variability, is computed by

$$\mathrm{PSRF} = \sqrt{\frac{n-1}{n} + \frac{B}{nW}},$$

where B is the variance of the between-chain means for m chains, and W is the average of the m within-chain variances. As the number of iterations n increases, PSRF should be close to 1 if the m chains have converged to the target posterior distribution. A value of PSRF smaller than 1.1 indicates convergence of the Markov chain, while a high value greater than 1.1 indicates lack of convergence and needs longer iterative simulations. In this thesis, convergence will be assessed by examining parameter trace plots as well as by Geweke (1992) and Gelman et al. (1992) diagnostics. We also want to make sure that the Markov chain is able to explore the entire parameter space and moves between separate high probability regions, which is known as mixing. Mixing is related to the acceptance rate and can also be visually monitored by checking trace plots for individual parameters.

In addition, the samples, especially consecutive samples, generated from MCMC algorithms normally show within chain correlation because each sample is dependent on the previous one. This autocorrelation can be reduced using the process of thinning, by only keeping every k_{th} simulation draw (after burn-in) from the posterior distribution and discarding the remaining samples. However, the practice of thinning can result in a loss of information and reduce the precision of the MCMC algorithms, and is often inefficient and unnecessary (Link and Eaton, 2012). However, Gelman et al. (1995) suggested that thinning is useful when computer storage is a problem in dealing with a large number of parameters. Link and Eaton (2012) argued that thinning can reduce the autocorrelation between MCMC samples. After a Markov chain has converged, the effective sample size n_{eff}, which is the approximate number of independent samples for any parameter of interest from the MCMC chain, can be calculated by

$$n_{\text{eff}} = \frac{n}{1+2\sum_{t=1}^{\infty}\rho(t)},$$

where n is the total number of samples after discarding the burn-in period, and $\rho(t)$ is the autocorrelation between the samples in the Markov chain at lag t.

Once a Markov chain has converged, summary statistics such as the posterior mean, median or mode of the MCMC samples can be reported as the point estimate for each parameter in question from its posterior distribution. The posterior mean and median can be used when the posterior distribution is continuous, but the latter is more robust when there are extreme values (outliers) present. The posterior mode is often used when the posterior distribution is discrete. Uncertainty in point estimates can be measured by using interval estimates, which are known as credible intervals in Bayesian statistics. A credible intervals is often constructed using percentiles of the posterior samples. A $100(1-\alpha)\%$ credible interval is defined by $(100 \times \alpha/2)_{\text{th}}$ (the lower bound) and $(100 \times (1-\alpha/2))_{\text{th}}$ (the upper bound) percentiles of the MCMC samples. It means that the parameter will lie in the interval with posterior probability $(1-\alpha)$.

2.2 Generalised linear models

Simple linear model is the simplest regression model which estimates a linear relationship between the response variable and the covariate data. It takes the form of

$$\begin{aligned} Y_i &\sim \mathrm{N}\,(\mu_i, \sigma^2), \;\; i = 1, \ldots, n, \\ \mathbb{E}(Y_i) &= \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \end{aligned}$$

where $(Y_1, \ldots, Y_n)$ are independent and each response variable Y_i is assumed to be Gaussian distributed with mean μ_i and variance σ^2. $\boldsymbol{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ contains a vector of p known covariates relating to observation i and a 1 for the intercept term. $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a column vector of covariate regression parameters including an intercept term β_0. This simple linear model can be extended to a generalised linear model (GLM) (Nelder and Wedderburn, 1972), where the response $\boldsymbol{Y}$ can be a set of independent random variables from any distribution in the exponential family, such as the Gaussian, Poisson, Binomial and Gamma distributions, rather than simply being Gaussian distributed. The exponential family is a flexible class of statistical distributions which, for a response variable Y_i whose probability distribution function $f(\cdot)$ depends on the parameter θ_i, takes the form of

$$f(Y_i = y_i | \theta_i) = \exp\left[a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)\right], \tag{2.2}$$

where a, b, c, d are known functions and $b(\theta_i)$ is called the natural parameter. The distribution is said to be in canonical form if $a(y_i) = y_i$.

A generalised linear model describes the relationship between the response and the linear predictor as

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.3}$$

where $g(\cdot)$ is a monotone, differentiable function called the link function, which relates the mean $\mathbb{E}(Y_i) = \mu_i$ to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$. The choice of the link function depends on the type of the response data. Identity link function $g(\mu_i) = \mu_i$ is used for Gaussian data; Logit link function $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$ can be used for Binomial data; Log link function $g(\mu_i) = \ln(\mu_i)$ is often used for Poisson count data.

2.2.1 Generalised linear models for count data

In this thesis, all the modelling approaches are developed to model count data, such as counts of the numbers of hospital admissions. An appropriate and commonly used model to represent count data is the Poisson distribution. Suppose $Y_i \sim \text{Poisson}(\mu_i)$ and are independent for $i = 1, \ldots, n$, then the probability mass function is:

$$\begin{aligned} f(Y_i = y_i|\mu_i) &= \frac{\mu_i^{y_i} \exp(-\mu_i)}{y_i!} \\ &= \exp\left(y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\right), \end{aligned} \tag{2.4}$$

where $\mu_i = \mathbb{E}(Y_i)$. Hence the Poisson distribution is a member of the exponential family of distributions, and a generalised linear model can be used to model these count data. Because the response data from the Poisson distribution are non-negative, the natural log function is a suitable choice of the link function. The Poisson GLM model takes the general form of

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i), \;\; i = 1, \ldots, n, \\ \ln(\mu_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta}. \end{aligned} \tag{2.5}$$

A Poisson model assumes that the mean and variance are the same, that is $\mu_i = \mathbb{E}(Y_i) = \text{Var}(Y_i)$ in our case. However, in many cases, a Poisson-distributed random variable can have much greater variance than the mean. This is called overdispersion and is often addressed by using a quasi-Poisson or negative binomial model (McCullagh and Nelder, 1989). Both approaches estimate an extra dispersion parameter to account for the extra variance. The former approach assumes that the variance is a linear function of the mean, whereas the latter assumes that the variance is a quadratic function of the mean. When the data contain an excess of zero counts, a zero-inflated Poisson model (Lambert, 1992) should be used, which is a mixture of a Poisson count model and a logit model for predicting excess zeros.

In the frequentist framework, generalised linear models typically use the iteratively reweighted least squares algorithm to find the maximum likelihood estimator (Dobson and Barnett, 2018), while in the Bayesian framework, parameters are often assigned a prior distribution and inference is typically based on the posterior distribution using Markov chain Monte Carlo simulation, such as Gibbs sampling and the Metropolis-Hastings algorithm.

2.3 Model comparison

A number of methods have been proposed to evaluate the relative quality of different statistical models applied to the same data. Likelihood is a measure of how likely one will obtain the observed data given a model. Although increasing the number of parameters tends to increase the likelihood, it might result in an overfitting issue. Hence, the best model is supposed to explain the greatest amount of variation in the observed data using the fewest independent parameters. Therefore, model selection criteria which attempt to achieve the trade-off between maximizing the likelihood and minimizing the risk of overfitting are necessary.

2.3.1 Akaike Information Criterion

Akaike Information Criterion (AIC) (Akaike, 1974) is a common criterion for model selection. The basic AIC is computed as

$$\mathrm{AIC} = -2\ln(\hat{L}) + 2K, \tag{2.6}$$

where $\hat{L}$ is the maximum value of the model likelihood measuring the goodness of model fit, and K is the number of estimated model parameters which is a penalty term discouraging overfitting. Adding more parameters is likely to improve the goodness of fit ($\hat{L}$), but it also improves the penalty term (K). Therefore, AIC deals with the balance between the model fit and the model complexity. The model with the lowest AIC is preferred when comparing two or more models.

2.3.2 Bayesian Information Criterion

Bayesian Information Criterion (BIC) (Schwarz et al., 1978) is another model selection criterion computed as

$$BIC = -2\ln(\hat{L}) + K\ln(n), \tag{2.7}$$

where n is the number of data points. Again, the model with the lowest BIC should be preferred when comparing multiple models. BIC applies a much larger penalty to complex models with large n than the AIC, thus it favors simpler models than the AIC.

2.3.3 Deviance Information Criterion

Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is an alternative comparison method. It is a generalisation of AIC based on the model deviance, and particularly used in Bayesian model selection. Deviance measures the deviation of a considered model from the true model which is a perfect fit to the data. In the Bayesian framework, deviance is defined as

$$D(\boldsymbol{\theta}) = -2\ln(f(\boldsymbol{Y}|\boldsymbol{\theta})).$$

Since the deviance tends to decrease as the number of parameters increase, it needs to be penalized by the independent number of parameters. The model with the minimum DIC value should be preferred when comparing multiple models. DIC is computed as DIC $=$ $\bar{D} + p_d$, where $\bar{D}$ is the expectation of the posterior deviance, that is

$$\bar{D} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left(-2\ln f(\boldsymbol{Y}|\boldsymbol{\theta})\right). \tag{2.8}$$

p_d is the effective number of parameters measuring the complexity of a model. Consider a Markov chain with S posterior simulations $\{\boldsymbol{\theta}^s, s = 1, \ldots, S\}$ for the model parameters $\boldsymbol{\theta}$, $\bar{D}$ can be computed by

$$\bar{D} = -2 \times \frac{1}{S}\sum_{s=1}^{S} \ln f(\boldsymbol{Y}|\boldsymbol{\theta}^s),$$

and p_d can be computed by

$$p_d = 2\left(\ln f(\boldsymbol{Y}|\bar{\boldsymbol{\theta}}) - \frac{1}{S}\sum_{s=1}^{S} \ln f(\boldsymbol{Y}|\boldsymbol{\theta}^s)\right),$$

where $\bar{\boldsymbol{\theta}} = \frac{1}{S}\sum\limits_{s=1}^{S} \boldsymbol{\theta}^s$.

2.4 Spatial statistics

Spatial statistics is the analysis and modelling of data that consist of observations at different spatial locations. There are three main types of spatial data; geostatistical data, point process data and areal data. For a geostatistical process, the spatial data consist of observations measured at many fixed and precise locations, e.g. the concentration of air pollution measured at a set of monitor stations within a study region. For a point process, the locations of the observations themselves are the data of interest. An example would be the locations of an earthquake. For an areal unit process, the entire study region is split into a set of non-overlapping areal units such as census tracts, administrative zones, electoral wards or ZIP codes, and the areal data are aggregated at the level of spatial units and often available as summary measurements such as case counts or rates referenced to areal units. For example, the percentage of unemployed people or the hospital admission counts in each areal unit. Areal data usually do not expose the geographical location of the individual residences, therefore they are frequently used in public health studies due to patient confidentiality concerns. In this thesis, I restrict the attention to the modelling of areal unit count data in a disease context, which is known as disease mapping.

2.4.1 Disease mapping

Disease mapping is the field of statistical epidemiology focusing on estimating the spatial pattern in disease risk across a study region (e.g. a city or country), evaluating the evolution of disease risk over time, as well as detecting areas of high risks. Most disease mapping studies utilise aggregated data such as counts of disease incidence or mortality collected from non-overlapping areal units (e.g. census tracts or health board areas) that comprise the study region, because patient-level data cannot be made publicly available due to confidentiality issues.

Consider a study region $\boldsymbol{A}$ which is partitioned into n non-overlapping areal units $\boldsymbol{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$. The response data are collected for each areal unit and are denoted by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. In this thesis models are developed based on areal unit count data in a disease context, so here Y_i represents the counts of the numbers of hospital admissions with a primary diagnosis of a particular disease in areal unit $\mathcal{A}_i$. However, only modelling the disease risk based on disease counts $\boldsymbol{Y}$ is misleading because of the deficient consideration in population size and demographics, which would vary from area to area. For example, an area that has a high proportion of elderly people tends to have higher counts of heart disease. In order to overcome the effects of confounding factors such as age and sex, the expected disease counts in each areal unit, denoted by $\boldsymbol{E} = (E_1, \ldots, E_n)$, are constructed by indirect standardisation based on the population size, age, and sex structure within each areal unit. They are calculated by stratifying the population living in each areal unit into a number of non-overlapping strata according to their age and sex demographics (e.g. male 0-9, male 10-19, etc), and then the population in each stratum is multiplied by the disease rate in that stratum for the whole study region. The sum of these over all strata is the expected disease counts for each areal unit. The formula for computing E_i for areal unit $\mathcal{A}_i$ is given by

$$E_i = \sum_{j=1}^{m} n_{ij} \times r_j, \tag{2.9}$$

where m is the number of strata, n_{ij} is the population of area $\mathcal{A}_i$ in stratum j, and r_j is the overall disease rate in stratum j. The standardised incidence ratio (SIR) is an exploratory estimate of disease risk calculated by $\text{SIR}_i = \frac{Y_i}{E_i}$, which is the ratio of the observed to the expected disease counts for each areal unit. An SIR value greater than 1 means that there are more disease cases observed than expected in the areal unit, suggesting an increased level of disease risk compared to the average during the study period, while a value less than 1 indicates fewer observed disease counts than expected in the areal unit, which corresponds to a decreased level of disease risk. For example, an SIR of 1.2 represents a disease risk which is 20% higher than the average during the study period, and an SIR of 0.9 corresponds to a 10% decrease in disease risk compared to the average.

Disease risk varies in space and time and is often affected by risk inducing factors such as environmental exposures (e.g. contamination of water, air pollution), population behaviours (e.g. smoking, alcohol consumption) and poverty (or socio-economic deprivation) (McCartney, 2012). Health agencies routinely produce disease maps to graphically illustrate such differences in disease risk, by displaying raw disease rates (the SIR) for each of the areal units. These maps allow to quantify the spatial inequalities in ill health across the study region and identify high-risk areas, which provide guidance for policy makers in disease intervention and health resources allocation. However, when the disease of interest is rare or the study region has small populations, some areal units could have small values of the expected number of disease cases, therefore the $\text{SIR} = \frac{Y_i}{E_i}$ would be extreme and unstable due to small random fluctuations in Y_i (Elliot et al., 2000), leading to unstable and uninformative disease risk estimates. Furthermore, visually examining maps of the SIR does not allow for the systematic identification of clusters of areas that exhibit high risks, which health agencies will want to target for risk reduction strategies. In addition, since data are collected over space we would expect that spatial correlation exists between areas that are spatially close to each other. Therefore, naively using the SIR ignores the potential spatial autocorrelation present in the data, and also does not consider the important covariate risk factors which could affect the data. In order to address these issues, it is necessary to develop model-based approaches which can capture the spatial variation in disease risk, separate the variation from random noise and account for the spatial autocorrelation structure in the data. Bayesian hierarchical models are commonly adopted to estimate the disease risk by using a combination of the covariates and a set of spatially varying random effects which account for any residual spatial autocorrelation after adjusting for covariates. These spatial random effects are typically modelled by using conditional autoregressive (CAR) priors, which will be discussed in Section 2.4.4.

2.4.2 Neighbourhood matrix

One feature of modelling the spatial data is that the spatial dependence in the data needs to be considered. The spatial autocorrelation structure for the n areal units is represented by a non-negative $n \times n$ symmetric neighbourhood or adjacency matrix $\boldsymbol{W}$, which specifies the spatial closeness between pairs of areal units. The elements $\{w_{ij}\}$ of the neighbourhood matrix $\boldsymbol{W}$ can be either continuous or binary, and in both cases a larger value of w_{ij} represents that areas $(\mathcal{A}_i, \mathcal{A}_j)$ are spatially closer to each other. A continuous $\boldsymbol{W}$ matrix is often based on distance and an example of this is $w_{ij} = \frac{1}{d_{ij}}$, where d_{ij} is the distance between centroids of areas $(\mathcal{A}_i, \mathcal{A}_j)$. However, this continuous specification leads to a dense $\boldsymbol{W}$ matrix, which can increase the computational intensity in the model fitting. Consequently, a binary neighbourhood matrix $\boldsymbol{W}$ is typically adopted, so that $w_{ij} = 1$ if areas $(\mathcal{A}_i, \mathcal{A}_j)$ are spatially close and $w_{ij} = 0$ otherwise. Commonly, the border sharing specification is used in the literature to determine $\boldsymbol{W}$, where $w_{ij} = 1$ if areas $(\mathcal{A}_i, \mathcal{A}_j)$ share a common geographical border (denoted $i \sim j$) and $w_{ij} = 0$ otherwise. As an area cannot be the neighbour of itself, $w_{ii} = 0$ for all i. Using this specification $\boldsymbol{W}$ is a sparse matrix, which makes the fitting of models more efficient. In this thesis, I refer to the neighbourhood matrix constructed from the border sharing rule defined above as the border sharing $\boldsymbol{W}$.

2.4.3 Moran's I test

Moran's I (Moran, 1950) is a common statistic used to measure the strength of spatial autocorrelation within a set of areal data. The level of spatial autocorrelation in $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is given as

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (Y_i - \bar{Y})^2},$$

where $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. The value of Moran's I ranges from -1 and 1. A value close to 1 indicates strong positive spatial autocorrelation among the data, a value close to -1 indicates strong negative spatial autocorrelation, and a zero value corresponds to spatial randomness and no autocorrelation. In real life applications, Moran's I statistic is mostly positive because areas which are close together in space are generally more likely to be spatially correlated and have similar data values. A permutation test can be conducted to test the null hypothesis that no spatial autocorrelation exists in the data, and the alternative hypothesis is that there is some spatial autocorrelation. The permutation test steps are

1. Calculate the observed Moran's I test statistic, I_{obs}.

2. Randomly permute the data K times and for each permuted data set calculate the Moran's I test statistic, denoted by $(I_1, \ldots, I_K)$.

3. Calculate the estimated two-sided p-value by

$$p = \frac{2}{K+1} \sum_{k=1}^{K} \mathrm{I}\left[I_k > |I_{obs}|\right],$$

where $\mathrm{I}[\cdot]$ denotes an indicator function, and is equal to 1 if $I_k > |I_{obs}|$ and is zero otherwise.

2.4.4 Spatial modelling

In this thesis the responses are areal unit count data in a disease context. Areal count data are generally modelled by extending the Poisson log-linear model (2.5) to account for the spatial autocorrelation in the data. Consider a study region partitioned into n non-overlapping areal units indexed by $i \in \{1, \ldots, n\}$. A response Y_i is observed in each of those areal units to give a set of response data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The expected number of disease cases for each areal unit are denoted by $\boldsymbol{E} = (E_1, \ldots, E_n)$, which can be computed via indirect standardisation. Covariate information, if relevant, is given by $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)$, where $\boldsymbol{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ is a row vector of p known covariates relating to areal unit i and a 1 for the intercept term. Due to the convenient structure of Bayesian hierarchical models, areal count data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are commonly modelled by extending the simple Poisson generalised linear model (2.5) to a generalised linear mixed model with a set of spatially correlated random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$. A general specification is given by

$$\begin{aligned} Y_i | E_i, R_i &\sim \text{Poisson}(E_i R_i), \quad i = 1, \ldots, n, \qquad (2.10) \\ \ln(R_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i, \end{aligned}$$

where R_i denotes the disease risk in areal unit i and is on the same scale as the SIR. The spatial random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ account for the spatial autocorrelation in the data and are typically modelled via a conditional autoregressive (CAR) prior, which can be specified by a set of univariate full conditional distributions of the form $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$ for $i = 1, \ldots, n$. The spatial autocorrelation between these random effects is typically accounted for by a binary neighbourhood matrix $\boldsymbol{W}$ based on the border sharing specification introduced in Section 2.4.2, where $w_{ij} = 1$ if areas (i, j) share a common geographical border and $w_{ij} = 0$ otherwise. Diagonal elements $w_{ii} = 0$ for all i. Many different CAR models have been proposed to model the spatial random effects $\boldsymbol{\phi}$, and four models that are the most popular are described below.

2.4.4.1 Intrinsic CAR model

The first and simplest CAR prior is the intrinsic model (Besag et al., 1991) and is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^2}{\sum_{j=1}^n w_{ij}}\right). \tag{2.11}$$

The conditional expectation of ϕ_i is the mean of the random effects in its neighbouring areal units, and thus each areal unit is modelled as being similar to its neighbours. The conditional variance is inversely proportional to the number of neighbouring units. This is sensible under the assumption of strong spatial autocorrelation in that the more neighbours an area has (i.e. $\sum_{j=1}^n w_{ij}$ increases), the more information there is in the data about the value of its random effect, as a result, the conditional variance goes down. This model has several limitations. The single variance parameter τ^2 does not determine the strength of the spatial correlation between the random effects. This prior is not appropriate when data are weakly correlated (Lee, 2011), because in such cases an increased number of neighbours would not necessarily result in more information about the random effect. Additionally, the intrinsic CAR prior is not appropriate for singleton areas that have no neighbours, because this will cause $\sum_{j=1}^n w_{ij} = 0$ and so lead to infinite mean and variance for ϕ_i. The joint probability distribution for $\boldsymbol{\phi}$ corresponds to an improper multivariate Gaussian distribution, which is given by

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

Here $\boldsymbol{Q}(\boldsymbol{W})^-$ denotes the Moore–Penrose (Moore, 1920, Penrose, 1955) generalised inverse of the singular precision matrix $\boldsymbol{Q}(\boldsymbol{W})$, where $\boldsymbol{Q}(\boldsymbol{W}) = \mathrm{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}$ and $\boldsymbol{W}\mathbf{1}$ is a vector containing the number of neighbours for each areal unit.

2.4.4.2 Convolution CAR model

Besag et al. (1991) also proposed the convolution (or BYM) CAR prior. It combines the intrinsic CAR model with a set of independent random effects, and is given by

$$\begin{aligned}\phi_i &= \phi_i^{(1)} + \phi_i^{(2)} \\ \phi_i^{(1)}|\boldsymbol{\phi}_{-i}^{(1)} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{ij}\phi_j^{(1)}}{\sum_{j=1}^n w_{ij}}, \frac{\tau_1^2}{\sum_{j=1}^n w_{ij}}\right) \\ \phi_i^{(2)} &\sim \mathrm{N}(0, \tau_2^2),\end{aligned} \tag{2.12}$$

where $\boldsymbol{\phi}$ now consists of two components. $\boldsymbol{\phi}^{(1)} = (\phi_1^{(1)}, \ldots, \phi_n^{(1)})$ is assigned the intrinsic CAR prior, and $\boldsymbol{\phi}^{(2)} = (\phi_1^{(2)}, \ldots, \phi_n^{(2)})$ is a set of independent and identically normally distributed random effects, with mean zero and common variance τ_2^2. The ratio of the two variance parameters $\frac{\tau_1^2}{\tau_2^2}$ controls the strength of spatial autocorrelation between random effects, which overcomes the limitation of the intrinsic CAR prior inducing too much spatial smoothness. A very large value of $\frac{\tau_1^2}{\tau_2^2}$ corresponds to strong spatial dependence between the random effects, whereas a very small value of $\frac{\tau_1^2}{\tau_2^2}$ corresponds to spatial independence. However, it is difficult to estimate $2 \times n$ random effects $\boldsymbol{\phi} = (\phi_1^{(1)}, \phi_1^{(2)}, \ldots, \phi_n^{(1)}, \phi_n^{(2)})$ given n data points, and only the sum $\phi_i^{(1)} + \phi_i^{(2)}$ is identifiable.

2.4.4.3 Proper CAR model

Stern and Cressie (2000) adapted the intrinsic model by adding a spatial autocorrelation parameter ρ, which allows the strength of spatial autocorrelation to be estimated from the data. The univariate full conditional distribution is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\rho\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^2}{\sum_{j=1}^n w_{ij}}\right). \tag{2.13}$$

Here ρ controls the level of spatial autocorrelation globally across the study region, with a value close to 1 corresponding to strong spatial dependence, while a value of zero corresponds to spatial independence. One problem with this model is that when $\rho = 0$ the random effects are assumed to be independent in space, however according to the formula, the conditional variance still depends on the number of neighbours that an area has. This issue was addressed by Leroux et al. (2000).

2.4.4.4 Leroux CAR model

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

$$\phi_i | \boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\rho \sum_{j=1}^{n} w_{ij}\phi_j}{\rho \sum_{j=1}^{n} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{n} w_{ij} + 1 - \rho}\right). \tag{2.14}$$

Again, $\rho \in [0,1]$ controls the strength of spatial autocorrelation in the data. If $\rho = 0$, the conditional mean is 0 and the conditional variance remains constant equal to τ^2. This suggests that the neighbouring random effects $\boldsymbol{\phi}_{-i}$ do not provide any information about ϕ_i, thus indicating independence in space. If $\rho = 1$, the model corresponds to the intrinsic model (2.11), indicating strong spatial autocorrelation between the random effects. The joint distribution for $\boldsymbol{\phi}$ is given by

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

where the precision matrix $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\mathrm{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$, which is an invertible matrix if $\rho \in [0,1)$. Here $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{I}$ is an $n \times n$ identity matrix. The partial correlation between (ϕ_i, ϕ_j) conditioning on the remaining spatial random effects (denoted $\boldsymbol{\phi}_{-ij}$) specified by this model is

$$\mathrm{Corr}\left(\phi_i, \phi_j | \boldsymbol{\phi}_{-ij}\right) = \frac{\rho w_{ij}}{\sqrt{\left(\rho \sum_{v=1}^{n} w_{iv} + 1 - \rho\right)\left(\rho \sum_{v=1}^{n} w_{jv} + 1 - \rho\right)}}. \tag{2.15}$$

Equation (2.15) shows that random effects (ϕ_i, ϕ_j) are modelled as partially correlated if $w_{ij} = 1$, otherwise, the partial correlation between (ϕ_i, ϕ_j) is 0 and they are modelled as conditionally independent. Hence the neighbourhood matrix $\boldsymbol{W}$ controls the spatial autocorrelation structure between the random effect terms. Thus, given the neighbourhood matrix $\boldsymbol{W}$ defined by the border sharing specification, if areas (i, j) share a common border $(w_{ij} = 1)$ their random effects are correlated and are smoothed over in the modelling process. Otherwise $(w_{ij} = 0)$ the random effects (ϕ_i, ϕ_j) are conditionally independent and are not smoothed towards each other. One disadvantage of the Leroux model is that it only has a single spatial dependence parameter ρ across the study region, assuming that the global level of dependence does not change over space.

2.5 Spatial-temporal modelling

The spatial models outlined in Section 2.4.4 are applied to data collected at a single time period across n areal units. This class of models has also been extended to the spatio-temporal domain in order to estimate the evolution of disease risk in both space and time. Now consider the areal count data that are collected for $t \in \{1,...,T\}$ consecutive time periods. The observed disease counts are denoted by $\boldsymbol{Y} = (\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n)$, where $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{iT})$ denotes the number of observed disease cases in areal unit i over all time periods. The expected number of disease cases are denoted by $\boldsymbol{E} = (\boldsymbol{E}_1,\ldots,\boldsymbol{E}_n)$, with $\boldsymbol{E}_i = (E_{i1},\ldots,E_{iT})$ denoting the expected disease counts in areal unit i over all time periods. Modelling such spatio-temporal data not only needs to account for spatial autocorrelation but also correlation in time. In this section, I introduce some of the important modelling approaches in the spatio-temporal disease mapping literature. Note that these models are described without including covariates, however the addition of covariate information is trivial.

2.5.1 Bernardinelli model

One of the earliest spatio-temporal models was proposed by Bernardinelli et al. (1995). The model allows for the analysis of risk using spatially correlated linear temporal trends for each areal unit. The disease risk R_{it} in areal unit i during time point t is modelled by

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n, \quad t=1,\ldots,T, \\ \ln(R_{it}) &= \mu + \phi_i + [\beta + \delta_i]t. \end{aligned} \tag{2.16}$$

Here μ is a global intercept term common to all areal units and β is an overall slope parameter for the linear time trend. Different areas are allowed to have different intercepts by introducing the random effects $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_n)$. Likewise, the linear slope can vary in space via $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)$. In other words, ϕ_i represents the difference between the area specific intercept for area i, $\mu+\phi_i$, and the overall intercept, μ. Similarly, δ_i indicates the difference between the trend slope for area i, $\beta+\delta_i$, and the global linear trend slope, β.

The random effects $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$ can be either spatially independent unstructured random effects, modelled by a Gaussian prior distribution with mean zero and constant variance, or spatially correlated structured random effects, modelled by the intrinsic CAR prior (Besag et al., 1991). The latter allows for the spatial autocorrelation in both intercepts and slopes.

Although $\delta_i \times t$ gives this model the increased flexibility of allowing for different temporal trends for different areal units, the time trend in disease variation is restricted to be linear which may not be appropriate for long time periods of data.

2.5.2 Knorr-Held model

Knorr-Held (2000) proposed an alternative spatio-temporal Bayesian model, which extends those models only with additively separable space and time main effects such as Knorr-Held and Besag (1998) by introducing an additional spatio-temporal interaction term. In this model, the response Y_{it} is assumed to have a binomial distribution, with parameters n_{it} and R_{it} being the number of people at risk and the binomial probability (disease risk) in areal unit i at time period t. The response data are modelled by a binomial generalised linear mixed model with a logit link function. The modelling framework is outlined as

$$\begin{aligned} Y_{it}|n_{it}, R_{it} &\sim \text{Binomial}(n_{it} R_{it}),\ i = 1, \ldots, n, \quad t = 1, \ldots, T, \\ \ln\left(\frac{R_{it}}{1 - R_{it}}\right) &= \mu + \theta_i + \phi_i + \alpha_t + \gamma_t + \delta_{it}, \end{aligned} \tag{2.17}$$

where μ is the overall intercept, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are area specific spatial effects, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$ are temporal effects. $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{nT})$ are spatio-temporal interaction terms which account for the disease variation that cannot be attributed to the separate space and time main effects. Here $\boldsymbol{\theta}$ are structured spatial random effects modelled by a conditional autoregressive (CAR) prior, and $\boldsymbol{\alpha}$ are structured temporal random effects modelled by a prior where neighbouring time points tend to have similar effects, e.g. a Gaussian first order random walk process given by $\alpha_1 \sim \text{N}(0, \sigma_\alpha^2)$ and $\alpha_t \sim \text{N}(\alpha_{t-1}, \sigma_\alpha^2)$ for $t = 2, \ldots, T$. $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are unstructured independent effects respectively in space and time and can be modelled using a Gaussian prior. Since both the spatial and temporal effects are divided into structured and unstructured components, this model can be thought as the extension of the convolution model outlined in Section 2.4.4.2 to the space-time domain.

Four different types of prior distribution are specified for the space-time interaction term δ_{it}, with each type corresponding to a different degree of autocorrelation in space and time. The first type is the interaction between the two unstructured space and time main effects, $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$. This is suitable when the variation explained by $\boldsymbol{\delta}$ does not contain any spatial or temporal structure, thus all interaction terms are independent. The second type is the interaction between the structured space effects $\boldsymbol{\theta}$ and unstructured time effects $\boldsymbol{\gamma}$, indicating that $\boldsymbol{\delta}$ are dependent over space but independent over time and can be modelled by CAR models. For an interaction between the unstructured $\boldsymbol{\phi}$ and structured $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ are spatially independent but dependent over time, and each areal unit follows a random walk process independent of other areas. Finally, for an interaction between the two structured space and time effects, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ are correlated in both space and time and can be modelled as $\delta_{it}|\boldsymbol{\delta}_{-it} \sim \mathrm{N}(\mu_{it}, v_{it})$. The conditional mean is computed as follows:

$$\mu_{it} = \begin{cases} \frac{1}{2}(\delta_{i,t-1} + \delta_{i,t+1}) + \frac{\sum_{j=1}^{n} w_{ij}\delta_{it}}{\sum_{j=1}^{n} w_{ij}} - \frac{\sum_{j=1}^{n} w_{ij}(\delta_{i,t-1}+\delta_{i,t+1})}{2\sum_{j=1}^{n} w_{ij}}, & \text{if } t \neq T \,\&\, t \neq 1, \\ \delta_{i,t+1} + \frac{\sum_{j=1}^{n} w_{ij}\delta_{it}}{\sum_{j=1}^{n} w_{ij}} - \frac{\sum_{j=1}^{n} w_{ij}\delta_{i,t+1}}{\sum_{j=1}^{n} w_{ij}}, & \text{if } t = 1, \\ \delta_{i,t-1} + \frac{\sum_{j=1}^{n} w_{ij}\delta_{it}}{\sum_{j=1}^{n} w_{ij}} - \frac{\sum_{j=1}^{n} w_{ij}\delta_{i,t-1}}{\sum_{j=1}^{n} w_{ij}}, & \text{if } t = T, \end{cases}$$

and the conditional variance is

$$v_{it} = \begin{cases} \frac{1}{\sum_{j=1}^{n} w_{ij}}, & \text{if } t = 1 \text{ or } t = T, \\ \frac{1}{2\sum_{j=1}^{n} w_{ij}}, & \text{if } t \neq 1 \,\&\, t \neq T. \end{cases}$$

This interaction term assumes that neighbouring areas tend to have similar temporal trends, and for each area spatial patterns near in time also tend to be similar. One advantage of this model is that it relieves the restrictive linear temporal trend assumption in Bernardinelli et al. (1995). Secondly, the model can be simplified by removing the spatio-temporal interaction once it turns out to be negligible. However, the increased number of parameters (five parameters for each data point) leads to more computational burden in the modelling process.

2.5.3 MacNab and Dean model

MacNab and Dean (2001) proposed a generalised additive mixed model for estimating disease risk by combining a conditional autoregressive (CAR) model for the spatial pattern and B-Splines smoothing (De Boor, 1972) for the temporal trend. Compared to Bernardinelli et al. (1995) and Knorr-Held (2000) who use a linear trend or random walk process to model the time trend, this model increases the flexibility in capturing a more complex temporal trend by using B-Spline smoothers. The model is of the form:

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n, \quad t=1,\ldots,T, \qquad (2.18)\\ \ln(R_{it}) &= \mu+\phi_i+S_0(t)+S_i(t), \end{aligned}$$

where μ is the overall intercept, ϕ_i is the spatial random effect for areal unit i, $S_0(t)$ is a fixed temporal effect common to all areal units, and $S_i(t)$ is an area specific temporal effect. The spatial effects $\boldsymbol{\phi}$ are modelled by a CAR prior and the global temporal trend $S_0(t)$ is modelled using a cubic B-Splines smoother. Two specifications are considered for the spatio-temporal interaction term $S_i(t)$. One approach models $S_i(t)$ linearly using $S_i(t) = S_i t$, and the other models $S_i(t)$ using cubic B-splines for each areal unit.

Modelling $S_i(t)$ using a random linear temporal trend is much simpler than using B-Splines smoothing which requires a large number of parameters to be estimated. Besides, the linear expression $\phi_i + S_i t$ provides a simple interpretation for the model, where ϕ_i is the spatial random effect and S_i represents the localised linear temporal trend above the overall mean trend. However, using B-Splines smoothers for both $S_0(t)$ and $S_i(t)$ may provide a more appropriate fit to data when longer time periods are considered. Alternative non-parametric smoothing approaches to estimate the time trend include MacNab and Gustafson (2007) and Torabi and Rosychuk (2011) which are B-Splines based models, and Ugarte et al. (2010) which uses P-Splines smoother.

2.5.4 Ugarte model

Ugarte et al. (2012) proposed a simplified model of that proposed by Knorr-Held (2000) by removing the two sets of unstructured spatial and temporal effects. Therefore, the model only contains three structured components and takes the form

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n, \quad t=1,\ldots,T, \qquad (2.19)\\ \ln(R_{it}) &= \mu+\phi_i+\alpha_t+\theta_{it}, \end{aligned}$$

where μ is the overall intercept, ϕ_i denotes the spatial random effect in areal unit i, α_t denotes the temporal effect during time t and θ_{it} represents a spatio-temporal interaction effect. The random effects $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ are respectively modelled by a Leroux CAR prior (Leroux et al., 2000) and a first order random walk process, and the interaction term θ_{it} is structured by a Gaussian distribution with mean zero and precision matrix being the Kronecker product of the precision matrices for the two main effects, $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$. Compared to that proposed by Knorr-Held (2000), this model is simpler because it has fewer model parameters to be estimated, however, since the temporal effect does not contain a parameter to determine the strength of the temporal correlation, the model can only be used for modelling strong temporal correlation and is not appropriate if the data are temporally weakly correlated.

2.5.5 Rushworth model

Rushworth et al. (2014) proposed to account for the spatio-temporal autocorrelation using a single set of non-separable random effects in space and time. The model is outlined as

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n, \quad t=1,\ldots,T, \\ \ln(R_{it}) &= \phi_{it}, \end{aligned} \tag{2.20}$$

where ϕ_{it} is the spatio-temporal random effect for areal unit i at time point t. The random effects at time point one, $\boldsymbol{\phi}_1 = (\phi_{11},\ldots,\phi_{n1})$, are specified using the Leroux CAR prior (Leroux et al., 2000) and have $\boldsymbol{\phi}_1 \sim \text{N}\left(\mathbf{0}, \tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1}\right)$, thus the spatial autocorrelation is induced through the precision matrix $\boldsymbol{Q}(\rho,\boldsymbol{W})$= $\rho(\text{diag}(\boldsymbol{W}\mathbf{1})-\boldsymbol{W})+(1-\rho)\boldsymbol{I}$. Temporal autocorrelation is induced amongst all other random effects via an autoregressive process of order 1, which is $\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} \sim \text{N}(\alpha\boldsymbol{\phi}_{t-1}, \tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1})$. Therefore, the random effects $\boldsymbol{\phi}_t$, except $\boldsymbol{\phi}_1$, are only dependent on the random effects $\boldsymbol{\phi}_{t-1}$. α controls the level of temporal autocorrelation in the data, which relieves the restriction of strong temporal dependence in the model proposed by Ugarte et al. (2012). This model only uses one set of space-time random effects to capture the spatio-temporal autocorrelation, thus only one variance parameter has to be estimated. However, the overall trends in the data can not be captured due to the absence of separate space and time random effects in the model.

2.5.6 Napier model

Napier et al. (2016) proposed an alternative approach to configure the spatio-temporal structure, by using an overall temporal trend and separate independent spatial effects for each time period. A Poisson log-linear variant of the model is given by

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n, \quad t=1,\ldots,T, \\ \ln(R_{it}) &= \phi_{it}+\theta_t, \\ \phi_{it}|\boldsymbol{\phi}_{-i,t},\boldsymbol{W} &\sim \text{N}\left(\frac{\rho_s\sum_{j=1}^n w_{ij}\phi_{jt}}{\rho_s\sum_{j=1}^n w_{ij}+1-\rho_s}, \frac{\tau_t^2}{\rho_s\sum_{j=1}^n w_{ij}+1-\rho_s}\right), \\ \theta_t|\boldsymbol{\theta}_{-t},\boldsymbol{D} &\sim \text{N}\left(\frac{\rho_\theta\sum_{j=1}^T d_{tj}\theta_j}{\rho_\theta\sum_{j=1}^T d_{tj}+1-\rho_\theta}, \frac{\tau_\theta^2}{\rho_\theta\sum_{j=1}^T d_{tj}+1-\rho_\theta}\right), \end{aligned} \tag{2.21}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_T)$ is an overall temporal effect common to all areal units, and $\boldsymbol{\phi}_t=(\phi_{1t},\ldots,\phi_{nt})$ denotes a separate spatial risk surface at each time period t. Each spatial surface $\boldsymbol{\phi}_t$ is modelled by the Leroux CAR prior (Leroux et al., 2000), and the corresponding joint multivariate Gaussian distribution is $\boldsymbol{\phi}_t \sim \text{N}\left(\mathbf{0},\tau_t^2\boldsymbol{Q}(\rho_s,\boldsymbol{W})^{-1}\right)$, where ρ_s is a spatial autocorrelation parameter and τ_t^2 is a temporally-varying variance parameter. Spatial autocorrelation is induced into areal units via the commonly used border sharing neighbourhood matrix $\boldsymbol{W}$. θ_t is assigned a one dimension Leroux CAR prior with a temporal autocorrelation parameter ρ_θ and variance parameter τ_θ^2. Here $\boldsymbol{D}=\{d_{tj}\}$ is a binary $T\times T$ temporal neighbourhood matrix analogously defined as the border sharing $\boldsymbol{W}$, with $d_{tj}=1$ if $|t-j|=1$ and $d_{tj}=0$ otherwise. ρ_s and ρ_θ respectively control the strength of the spatial and temporal autocorrelation, with a value of 1 indicating strong dependence and a value of 0 indicating independence. The model has an advantage of allowing the amount of spatial variation in the data to change over time, rather than assuming an overall spatial risk surface common to all time periods as in Knorr-Held (2000). This model can be implemented in the CARBayesST package (Lee et al., 2018) in R programming.

2.6 Identification of risk discontinuities

There are two primary goals in disease mapping studies (Waller and Carlin, 2010), namely: (a) providing accurate local disease risk estimates for each area and (b) detecting high/low-risk areas. Models incorporating spatial random effects with conditional autoregressive (CAR) priors (see Section 2.4.4) have been the mainstream to achieve the former goal, by smoothing the disease risks in neighbouring areas towards each other to remove random noise. In these models, the spatial autocorrelation structure in the data is fixed and typically induced by the neighbourhood matrix $\boldsymbol{W}$ based on the border sharing specification. Therefore these CAR models assume that there is a constant level of spatial smoothness across the entire study region, in other words, neighbouring areas are forced to have similar risks and so the spatial pattern in disease risk is modelled to be globally smooth. However, this assumption is somehow contradictory to the second goal of detecting high-risk areas. In practice, the level of spatial autocorrelation may vary across the study region and the risk surface may exhibit localised spatial autocorrelation structure. Some pairs of neighbouring areas are likely to be independent of each other and exhibit significantly different disease risks. These abrupt changes in disease risk can be driven by complex reasons such as the social, economic or environmental characteristics of adjacent neighbourhoods (Mitchell and Lee, 2014). Hence an excess of smoothing may blur or even conceal any discontinuities in the risk surface, prevent the identification of areas with elevated risks and lead to biased risk estimates. A variety of methods concerned with the identification of discontinuities in the risk surface have been developed in parallel to smoothing methods, including the fields of spatial clustering and boundary analysis. Spatial clustering approaches allow for discontinuities by identifying local spatially contiguous or non-contiguous clusters of areas that have elevated or reduced risks compared to their neighbours, while boundary analysis approaches allow for discontinuities by detecting the locations of risk boundaries that separate pairs of neighbouring areas of higher and lower disease risks.

2.6.1 Spatial clustering

In this section I introduce some statistical techniques which account for spatial discontinuities in the disease risk pattern by identifying spatial clusters of areas that exhibit substantially different risks (higher or lower risk) compared to their neighbours. These methods produce “closed boundaries” which entirely enclose a group of areal units with similar risks. One of the first approaches is scan statistics (Kulldorff, 1997), which can be implemented with the SaTScan software. It takes the form of a maximum likelihood ratio test based on the observed and expected cases inside and outside a potential cluster based on a search window with a certain shape. However, scan statistics only identify high-risk clusters and is unable to estimate the disease risk at the same time. Therefore, a number of Bayesian hierarchical models have been developed.

Knorr-Held and Raßer (2000) proposed to partition all areal units into a set of spatially contiguous clusters and the risk within each cluster is assumed to be constant. In the model, a set of areal units are selected as cluster centers and each of the remaining areal units is assigned to the cluster which has the minimal distance between the cluster center and the areal unit, where the distance is defined as the minimal boundaries that have to be crossed for moving from one to the other unit. This model has its adaptive nature in that the number of clusters, the location of clusters, the cluster memberships and the constant relative risk in each cluster are unknown parameters and estimated by the data, which sharply contrasts to CAR models in Section 2.4.4, where the number of parameters are known and the spatial smoothing is enforced across all neighbouring areas. However, inference for this model involves the complex and computationally expensive reversible jump Markov chain Monte Carlo algorithm (Green, 1995), which is likely to have multimodal problems. Charras-Garrido et al. (2012) developed a risk partition model for mapping the disease risk classes (or clusters) based on a discrete hidden Markov random field (HMRF) model. The observed data are augmented by a set of hidden variables that represent which risk class each areal unit belongs to. All the classes are naturally ordered by their risk levels and each areal unit is assigned to one of these classes, with a penalty term penalizing the neighbouring areas according to their distance between the risk classes. Thus, neighbouring areas are more likely to be correlated and have similar disease risk if their risk classes are closer. Parameters are estimated using a Monte Carlo Expectation-Maximisation algorithm, and the classification is carried out using a post-processing step.

Wakefield and Kim (2013) developed a Bayesian version of the Kulldorff (1997) approach. It specifies the number of clusters in advance and creates a list of candidate clusters by taking each areal unit in turn and sequentially adding the geographically closest neighbouring area (in terms of the distance to its centroid) until a pre-specified maximum cluster size is reached. The significance of clustering and which clusters have a high (or low) disease risk are evaluated through a number of posterior summaries on the number of clusters and cluster configurations. The approach rectifies a number of drawbacks of the scan statistic approach, such as a *p*-value threshold for significance has to be specified to test the presence of clusters, however, the approach is restricted by the requirement for circular clusters.

Anderson et al. (2014) proposed a two-stage procedure to model discontinuities in the spatial risk pattern. In the first stage, a set of candidate cluster configurations are elicited by applying a spatially-adjusted hierarchical agglomerative clustering algorithm to data. In the second stage, a Bayesian hierarchical model which includes cluster fixed effects and a CAR spatial random effect is fitted to all cluster configurations, and finally the cluster configuration leading to the minimum Deviance Information Criterion (DIC) is selected.

This approach treats the estimation of the cluster structure as a model comparison problem, which is straightforward and easy to implement. Additionally, the combination of a CAR model with a piecewise constant cluster model allows similar but not identical disease risks within one cluster and different risks between clusters, rather than naively assuming constant disease risk within a cluster as in Knorr-Held and Raßer (2000), Charras-Garrido et al. (2012) and Wakefield and Kim (2013), which may not be realistic in practice. Adin et al. (2019) extended the approach of Anderson et al. (2014) by replacing the cluster fixed effects with random effects and considering a spatio-temporal rather than a spatial setting. However, both papers have the common computational limitation of fitting multiple Bayesian models separately, as well as ignoring the uncertainty about the number of clusters in the data. Anderson et al. (2016) addressed these deficiencies in a purely spatial setting, by estimating disease risk and the cluster structure simultaneously in a single model, thus quantifying the uncertainty in the estimated cluster structure. They do this by first using an agglomerative hierarchical clustering to create a set of candidate cluster structures, each of which corresponds to a candidate neighbourhood matrix. Then they fit a single model that treats $\boldsymbol{W}$ as a parameter to be estimated, and assign it a discrete uniform prior whose values are the set of candidate neighbourhood matrices previously constructed. However, the clustering of this approach relies on a single clustering algorithm, which means that for each fixed number of clusters only one spatial cluster structure is allowed as a candidate in the model, thus the true cluster structure may not be identified in stage one. Anderson et al. (2017) proposed a model to identify clusters of areal units which are similar in terms of average disease risk and temporal trends during the study period. This model is an extension of that proposed by Bernardinelli et al. (1995) by introducing two other sets of parameters, which are cluster-specific intercept terms and cluster-specific linear trend slopes. Therefore, areas within the same spatial cluster would have similar disease risks and areas within the same temporal cluster would have similar temporal trends. The proposed model comes with both strengths and weaknesses. It improves the Bernardinelli et al. (1995) model by including a clustering mechanism, where a different intercept and linear slope is assigned to each cluster. In addition, it allows disease risk to vary within one cluster via the area-specific random effects, which more accords with real data. However, the number of clusters is required to be defined in advance, rather than being estimated from the data. Santafé et al. (2021) proposed to first estimate a single cluster configuration using a density-based clustering algorithm, and then a Bayesian hierarchical spatial model that takes the cluster configuration into account is fitted. In contrast to the previous proposals in Anderson et al. (2014) and Adin et al. (2019), this approach is able to automatically detect the number of spatial clusters, and is suitable for analysing large spatial data sets because only one disease mapping model is fitted based on the single cluster structure. However, the uncertainty in the cluster structure is not quantified, and an accurate cluster structure is required in the first step in order to provide precise risk estimates.

In Chapters 3, 4 and 5 of this thesis, I will focus on developing methodologies for simultaneously estimating the spatial patterns in disease risk and identifying clusters of areas that exhibited elevated or reduced risks compared to their geographical neighbours in the Bayesian disease mapping context.

2.6.2 Boundary analysis

The other class of statistical techniques accounts for spatial discontinuities in disease risk by identifying the locations of boundaries in the spatial surface that separate two geographically adjacent areas exhibiting vastly different risks. Thus the identified boundaries do not necessarily enclose an areal unit or group of units, which are known as "open boundaries".

Lu and Carlin (2005) proposed to calculate the absolute differences in the estimated disease risk between all pairs of neighbouring areas, which are called boundary likelihood values (BLVs). The boundaries are detected based upon the posterior distribution of the BLVs computed using Markov chain Monte Carlo simulation in a Bayesian hierarchical model. The border between each pair of adjacent areas is identified as a boundary if the posterior mean of the BLVs is greater than a certain pre-selected cutoff c_1. Alternatively, we can make probability statements by calculating the posterior exceedance probability of BLVs exceeding some cutoff. An advantage of this model is that it allows for direct probability statements regarding the likelihood that a geographical border is a boundary. However, the cutoff value needs to be specified by the user, indicating that the number of boundaries identified is essentially artificially determined in advance. Li et al. (2011) took a different approach, by treating the boundary detection as a model comparison problem. They fit a class of models with different neighbourhood matrices applied. These neighbourhood matrices represent a set of neighbourhood structures with different potential boundaries, and the one leading to the smallest Bayesian Information Criterion (BIC) is chosen.

A number of models estimate the local boundaries by treating the elements $\{w_{ij}\}$ of the neighbourhood matrix $\boldsymbol{W}$ as a set of binary random variables taking values 0 or 1 if areas (i, j) share a common border, rather than fixing them at 1. Therefore a boundary is said to exist between the two adjacent areas (i, j) when estimating $w_{ij} = 0$, which implies that the random effects (ϕ_i, ϕ_j) are conditionally independent and should not be smoothed over. If estimating $w_{ij} = 1$ then there is no boundary between areas (i, j), hence the two random effects are correlated and should be smoothed towards each other (see equation (2.15)). One of the first such approaches was proposed by Lu et al. (2007), who modelled elements $\{w_{ij}|i \sim j\}$ through a logistic regression model using a set of covariates which measure the dissimilarity between areas i and j. The model has the advantage of estimating the degree of spatial smoothing using both data and the related covariate information. However, a large collection of sensible covariates would be required in order to define a rich enough spatial weights matrix. In addition, the number of parameters in the logistic regression will increase quadratically as a function of the number of spatial units in the study. This could involve computationally expensive MCMC algorithms and cause problems of parameter identifiability, which also arise in a similar model proposed by Ma et al. (2010). Lee and Mitchell (2012) proposed an approach for capturing localized spatial autocorrelation structure by treating the set of $\{w_{ij}|i \sim j\}$ as a deterministic function of a small number of regression parameters and measures of dissimilarity based on covariate information. This model has the advantages of being fully automatic and parsimonious over the existing models. The main drawback is that the approach is crucially dependent on the existence of good quality dissimilarity measures. This issue has been addressed by the same authors in Lee and Mitchell (2013). They proposed an iterative algorithm which cycles between updating $\boldsymbol{W}$ and the remaining model parameters $\boldsymbol{\Theta}$ conditional on each other until a convergence criterion is met. Conditional on $\boldsymbol{W}$, the estimation of $\boldsymbol{\Theta}$ is fully Bayesian, whereas the elements $\{w_{ij}|i \sim j\}$ are treated as hyperparameters which are updated deterministically based on the current marginal posterior distributions of the random effects ϕ_i and ϕ_j. If the two marginal 95% posterior credible intervals for (ϕ_i, ϕ_j) overlap then we set $w_{ij} = 1$, otherwise we set $w_{ij} = 0$ and a boundary is found between areas (i, j). The elements $\{w_{ij}|i \sim j\}$ are not updated in a fully Bayesian setting, because only the estimates are provided rather than full posterior distributions, thus this approach cannot quantify the level of uncertainty in w_{ij}. Rushworth et al. (2017) proposed to treat the elements $\{w_{ij}|i \sim j\}$ as parameters with support on the unit interval to allow adaptive levels of spatial smoothing. Lee et al. (2021) developed a graph-based optimisation algorithm for estimating either static or temporally evolving risk boundaries. The algorithm views the areal units as the vertices of a graph and the neighbour relations as the set of edges. Firstly, the algorithm is applied to the data to estimate which edges in the graph should be removed via an objective function. An appropriate neighbourhood matrix $\boldsymbol{W}$ is then estimated by setting $w_{ij} = 0$ if the edge between areas (i, j) is removed, which suggests a boundary between the two areas. Otherwise, $w_{ij} = 1$ and no boundary exists between the two areas. Secondly, a Poisson log-linear model with spatio-temporally correlated random effects is fitted using the estimated $\boldsymbol{W}$. However this algorithm operates via a local search method and is not guaranteed to find the global optimal neighbourhood matrix. In addition, the approach also cannot measure the uncertainty in $\boldsymbol{W}$ when estimating disease risk.

In Chapter 6 of this thesis, I will propose an approach for jointly estimating disease risk and identifying the locations of boundaries that correspond to sizeable changes in disease risk between geographically adjacent areas.

2.7 Clustering algorithms

In this section, I will introduce a range of classical clustering methods which will be used in the methodology developed in Chapters 3, 4 and 5.

Initial interests in cluster analysis began in the 1960s, and the first application of clustering was in the disciplines of biology and ecology (Sneath et al., 1973). Clustering is the unsupervised classification of a set of unlabeled objects into groups called clusters according to their similarities on some specified characteristic(s). It has been broadly used in many fields such as data mining, image analysis, biology, business etc. (Madhulatha, 2012). For example, clustering has been applied to identify groups of genes that have similar functions. In marketing research, clustering can be used to separate customers into different clusters of people with different consumption habits. In meteorology, clustering has been used to find patterns in the atmosphere pressure of polar regions. A cluster is a collection of objects which are more similar to each other than they are to those belonging to other clusters. The similarity (or dissimilarity) between objects or clusters are normally defined by a distance measure. Some examples of usual distance functions are described in Section 2.7.1, including the Euclidean distance, Manhattan distance, Chebychev distance and Minkowski distance. A number of clustering approaches using different algorithms to constitute a cluster have been proposed. Section 2.7.2 discusses some popular clustering algorithms which will be used in this thesis.

2.7.1 Dissimilarity metrics

Consider a set of n objects which are described by data $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, where $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})$ is a data vector of length p which contains the information relating to object i. The dissimilarity (or distance) between objects i and j is denoted by d_{ij}. Four commonly used distance measures are described below.

- **Euclidean distance:** It is the most common choice of dissimilarity measurement which computes the root sum-of-squares of differences between objects. The Euclidean distance between objects (i, j) is computed as

$$d_{ij} = \sqrt{\sum_{v=1}^{p} (\xi_{iv} - \xi_{jv})^2}.$$

- **Manhattan distance:** It computes the sum of absolute differences between objects, which is given by

$$d_{ij} = \sum_{v=1}^{p} |\xi_{iv} - \xi_{jv}|.$$

- **Chebyshev distance:** It computes the maximum absolute differences between objects, which is given by

$$d_{ij} = \max\left\{|\xi_{i1} - \xi_{j1}|, \ldots, |\xi_{ip} - \xi_{jp}|\right\}.$$

- **Minkowski distance:** It is a generalised metric distance. The Minkowski distance between objects (i, j) is defined as

$$d_{ij} = \left(\sum_{v=1}^{p} |\xi_{iv} - \xi_{jv}|^q \right)^{\frac{1}{q}}, \ q \geq 1.$$

  Here, $q = 1$ and $q = 2$ refer to the Manhattan and Euclidean distance respectively, and $q = \infty$ corresponds to the Chebyshev distance.

2.7.2 Clustering methods

Suppose n objects are partitioned into K clusters by a clustering method, and the resultant cluster structure is denoted by $\boldsymbol{C}_K = (C_K^1, \ldots, C_K^K)$, where C_K^j represents the set of objects belonging to the j_{th} cluster in $\boldsymbol{C}_K$.

2.7.2.1 K-means clustering

K-means clustering (MacQueen et al., 1967) assigns each object to its closet centroid (center) by the Euclidean distance, and the collection of objects assigned to the same centroid forms a cluster. A centroid is defined as the average of all the objects in the cluster. The goal of this method is to minimize the within-cluster sum-of-squares. The basic k-means clustering algorithm for generating a cluster structure with K clusters is outlined as follows.

Algorithm 1: Basic k-means clustering algorithm

```
1. Randomly select $K$ objects as initial centroids, denoted by $\boldsymbol{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K)$;
repeat
    2. Assign each object to its nearest centroid based on the Euclidean distance, which
       is $d_{ij} = \sqrt{\sum_{v=1}^{p} (\xi_{iv} - O_{jv})^2}$ between object $i$ and centroid $\boldsymbol{O}_j$;
    3. Recalculate the centroid for each cluster;
until Centroids do not change.
```

2.7.2.2 K-medoids clustering

In contrast to k-means clustering, k-medoids clustering (Park and Jun, 2009) takes the object whose average dissimilarity to all the other objects in the cluster is minimal as the cluster center. k-medoids clustering is more robust than k-means clustering as a medoid is less influenced by outliers or extreme values than a mean (Madhulatha, 2011). Partitioning around medoids (PAM) algorithm is one of the earliest realisation of k-medoids clustering, which is outlined as follows.

Algorithm 2: Basic k-medoids clustering algorithm

```
1. Randomly select $K$ objects as initial medoids, denoted by $\boldsymbol{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K)$;
repeat
    2. Assign each object to its closet medoid. The Euclidean distance between object $i$
    and medoid $\boldsymbol{O}_j$ is $d_{ij} = \sqrt{\sum_{v=1}^{p} (\xi_{iv} - O_{jv})^2}$;
    3. Recalculate the medoid for each cluster. For each object $i$ in cluster $C_K^j$, compute
    its average dissimilarity to all the other objects in that cluster as
    $\sqrt{\frac{1}{N_j - 1} \sum_{f \in C_K^j} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_f)^2}$, where $N_j$ is the number of objects in cluster $C_K^j$. Then
    select the object with the minimum average dissimilarity as the new medoid;
until Medoids do not change.
```

2.7.2.3 Hierarchical agglomerative clustering

Hierarchical agglomerative algorithm (Hastie et al., 2009) begins with each object being in a separate cluster of its own and then iteratively merges the two least dissimilar clusters into a larger cluster until only one cluster containing all data points remains. The result of a hierarchical clustering algorithm is usually a hierarchical set of clusters represented by a tree diagram or dendrogram, and all the objects can be partitioned into a desired number of clusters by cutting the dendrogram at a given level. The dissimilarity D_{ij} between clusters C_K^i and C_K^j can be defined by various linkage methods, such as single linkage, centroid linkage, complete linkage, average linkage and Ward's linkage.

1. **Single linkage** (Florek et al., 1951, Sneath, 1957): It measures the dissimilarity as the minimum distance between any two objects from opposite clusters, which is computed as

$$D_{ij} = \min\left\{ ||\boldsymbol{\xi}_v - \boldsymbol{\xi}_w||, v \in C_K^i, w \in C_K^j \right\},$$

   where $||\cdot||$ denotes a distance metric. However, single linkage has the limitation of frequently suffering from the chaining effect and producing long straggly clusters that are difficult to interpret (Hartigan, 1981, Kuiper and Fisher, 1975).

2. **Centroid linkage**: It measures the dissimilarity as the distance between the centroids for two opposite clusters, which is computed as

$$D_{ij} = \left|\left| \frac{1}{N_i} \sum_{v \in C_K^i} \boldsymbol{\xi}_v - \frac{1}{N_j} \sum_{w \in C_K^j} \boldsymbol{\xi}_w \right|\right|,$$

where N_i and N_j are the number of objects in clusters C_K^i and C_K^j respectively.

3. **Complete linkage** (Sorensen, 1948): It measures the dissimilarity as the maximum distance between any two objects from opposite clusters, which is computed as

$$D_{ij} = \max\left\{||\boldsymbol{\xi}_v - \boldsymbol{\xi}_w||, v \in C_K^i, w \in C_K^j\right\}.$$

4. **Average linkage** (Sokal and Michener, 1958): It measures the dissimilarity as the average distance between all pairs of objects from opposite clusters, which is computed as

$$D_{ij} = \frac{1}{N_i \times N_j} \sum_{v \in C_K^i} \sum_{w \in C_K^j} ||\boldsymbol{\xi}_v - \boldsymbol{\xi}_w||.$$

5. **Ward's linkage** (Ward Jr, 1963, Murtagh and Legendre, 2014): It measures the dissimilarity as the increase in the total within-cluster sum-of-squares when joining two smaller clusters into a larger cluster, which is computed as

$$D_{ij} = SS\left(C_K^i \cup C_K^j\right) - \left[SS\left(C_K^i\right) + SS\left(C_K^j\right)\right],$$

where $SS(\cdot)$ denotes the within-cluster sum-of-squares, which is computed as $SS\left(C_K^i\right) = \sum_{v \in C_K^i} \left|\left|\boldsymbol{\xi}_v - \frac{1}{N_i} \sum_{v \in C_K^i} \boldsymbol{\xi}_v\right|\right|^2.$

Each linkage method defines the distance between two clusters in a unique way. The selected linkage will determine the way in which clusters are merged and so directly affect the clustering results.

2.7.2.4 Hierarchical divisive clustering

Hierarchical divisive algorithm begins with all objects in one cluster and at each step of iteration the most heterogeneous cluster is divided into two subclusters, which make up a so-called bipartition of the former cluster. This process is iterated until each object forms a singleton cluster. One of the divisive algorithms was proposed by Kaufman and Rousseeuw (2009). Two main choices should apply in the algorithm, which are outlined below.

1. **The choice of the cluster to be split:** At each iteration the cluster with the largest diameter is selected to be split. Here, the diameter of a cluster C_K^i is defined as the largest dissimilarity between all pairs of objects in the cluster, and is calculated as $\max\left\{||\boldsymbol{\xi}_v - \boldsymbol{\xi}_w||, v \in C_K^i, w \in C_K^i\right\}.$

2. **The bipartition of the selected cluster:** To divide the selected cluster C_K^i, its object which has the largest average dissimilarity to the other objects in C_K^i is selected to initiate one subcluster, denoted by C_{K+1}^l. Then each of the remaining objects in C_K^i is assigned to cluster C_{K+1}^l as long as its average dissimilarity to C_K^i is greater than that to C_{K+1}^l. What remained from the original cluster C_K^i naturally forms the other subcluster. The average dissimilarity from object $v \in C_K^i$ to the other objects in the same cluster is calculated as $\frac{1}{N_i-1} \sum_{w \in C_K^i, w \neq v} ||\boldsymbol{\xi}_v - \boldsymbol{\xi}_w||$.

Therefore, a new cluster structure $\boldsymbol{C}_{K+1}$ containing $K+1$ clusters is generated from the original cluster structure $\boldsymbol{C}_K$ with K clusters. The hierarchical divisive algorithm proposed by Kaufman and Rousseeuw (2009) can be implemented with the "cluster" package in R programming (R Core Team, 2013).

In k-means, k-medoids and hierarchical clustering methods, the number of clusters is chosen based on prior expert knowledge or plotting tools. For example, the elbow method plots the curve of the within-cluster sum of squares against varying number of clusters K. The location of an elbow point, after which the curve appears to level off, is generally considered as an indicator of the appropriate number of clusters. However, the elbow method would be subjective and ambiguous sometimes. Other objective and robust methods include the average silhouette method (Rousseeuw, 1987) and the gap statistic (Tibshirani et al., 2001). Average silhouette method computes the average silhouette statistics of objects for different values of K, and the value with the maximum average silhouette is the best choice. Gap statistic method compares the total within-cluster variation for different values of K with their expected values under the uniform reference distribution of the data. The gap statistic is estimated as the deviation of the total within-cluster variation from its expected value for each value of K, and the optimal K is the value that maximizes the gap statistic.

2.7.2.5 Model-based clustering

The clustering methods introduced above generate clusters based on the dissimilarity or distance between data points rather than probability models. An alternative approach is model-based clustering (Fraley and Raftery, 2002) which assumes that the data are generated by a mixture of underlying probability distributions. Each distribution corresponds to a different cluster and the goal is to estimate the parameters of these distributions. The parameters can be estimated using the expectation-maximisation (EM) algorithm (Dempster et al., 1977). A mixture of Gaussian distributions is considered in most occasions, and in this context the EM algorithm aims to estimate the mean and variance for each Gaussian distribution. The EM algorithm works by iteratively performing an E-step, which assigns each object to its most likely cluster, and an M-step, which then estimates the parameters by maximising the complete data log-likelihood based on the assigned objects. The algorithm ends up providing the probabilities that each object belongs to each cluster, and the final cluster of the object is determined by the maximum probability.

Consider a set of n objects, described by data $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, are from a mixture model with K clusters. The data likelihood is given by

$$L(\alpha_1, \ldots, \alpha_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) = \prod_{i=1}^{n} \sum_{j=1}^{K} \alpha_j f_j(\boldsymbol{\xi}_i | \boldsymbol{\theta}_j),$$

where $f_j(\boldsymbol{\xi}_i|\boldsymbol{\theta}_j)$ denotes the probability distribution for the j_{th} cluster, and $\boldsymbol{\theta}_j$ are the parameters of that distribution. $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ denotes the mixing proportions for each cluster with the constraints $\sum_{j=1}^{K} \alpha_j = 1$ and $\alpha_j > 0$. Let $\mathbf{z}_i = \{z_{i1}, \ldots, z_{iK}\}$ be the latent variables with each z_{ij} given as

$$z_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{\xi}_i \text{ comes from the } j_{th} \text{ cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose the probability density of $\boldsymbol{\xi}_i$ given $\mathbf{z}_i$ is specified as $\prod_{j=1}^{K} f_j(\boldsymbol{\xi}_i|\boldsymbol{\theta}_j)^{z_{ij}}$ and each $\mathbf{z}_i$ is independent and identically distributed from a multinomial distribution with probabilities $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, then the complete data log-likelihood is given by

$$l(\mathbf{z}_1, \ldots, \mathbf{z}_n; \alpha_1, \ldots, \alpha_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, |\boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log\left(\alpha_j f_j\left(\boldsymbol{\xi}_i|\boldsymbol{\theta}_j\right)\right).$$

In E-step, since the latent variable z_{ij} is unknown, the conditional expected value $\hat{z}_{ij}$ is substituted for z_{ij}. By Bayes' theorem, we have

$$\hat{z}_{ij} = \mathbb{E}\left(z_{ij}|\boldsymbol{\xi}_i, \alpha_1, \ldots, \alpha_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\right) = \frac{\hat{\alpha}_j f_j(\boldsymbol{\xi}_i|\hat{\boldsymbol{\theta}}_j)}{\sum_{s=1}^{K} \hat{\alpha}_s f_s(\boldsymbol{\xi}_i|\hat{\boldsymbol{\theta}}_s)}.$$

In M-step, the model parameters are estimated by maximising the log data likelihood given by

$$l(\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_n; \alpha_1, \ldots, \alpha_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{K} \hat{z}_{ij} \log\left(\hat{\alpha}_j f_j\left(\boldsymbol{\xi}_i|\boldsymbol{\theta}_j\right)\right).$$

Consider a Gaussian mixture model such that each cluster of objects follows a mulitvairate Gaussian distribution $f_j(\boldsymbol{\xi}_i|\boldsymbol{\theta}_j) \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The parameters $\boldsymbol{\theta}_j$ consist of a mean vector $\boldsymbol{\mu}_j$ and a covariance matrix $\boldsymbol{\Sigma}_j$ and are estimated by

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^n \hat{z}_{ij}\boldsymbol{\xi}_i}{\sum_{i=1}^n \hat{z}_{ij}},$$
$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^n \hat{z}_{ij}(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_j)^\top}{\sum_{i=1}^n \hat{z}_{ij}}.$$

The mixing proportions $(\alpha_1, \ldots, \alpha_K)$ are updated with

$$\hat{\alpha}_j = \frac{\sum_{i=1}^n \hat{z}_{ij}}{n}.$$

These two steps are repeated until convergence is reached. A usual criterion for convergence is that the difference in latent variables $\boldsymbol{z}_i$ between consecutive iterations is within a certain tolerance level. This model-based clustering algorithm can be implemented with the "mclust" package in R programming.

2.7.3 Adjusted rand index

One of the common statistics for measuring the clustering performance is the adjusted Rand Index (ARI) proposed by (Hubert and Arabie, 1985). It is a measure of the similarity between two cluster structures, with a larger value indicating a higher agreement between two cluster structures. A value of 1 indicates complete agreement between two cluster structures, a value of 0 indicates that the data points are randomly allocated to the two cluster structures, and a value less than 0 indicates that the level of agreement between two cluster structures is smaller than that if data points are randomly allocated.

Suppose n objects are partitioned into two different cluster structures, $\boldsymbol{C}_K = (C_K^1, \ldots, C_K^K)$ with K clusters and $\boldsymbol{S}_V = (S_V^1, \ldots, S_V^V)$ with V clusters. C_K^j and S_V^i respectively denote the set of objects in the j_{th} and i_{th} cluster in structures $\boldsymbol{C}_K$ and $\boldsymbol{S}_V$. Then the overlap between these two structures can be summarised in Table 2.1, where each entry n_{ji} denotes the number of objects in common between clusters C_K^j and S_V^i.

Table 2.1: A table summarising the number of objects in common between clusters C_K^j and S_V^i.

$\boldsymbol{S}_V$ $\boldsymbol{C}_K$	S_V^1	S_V^2	$\dots$	S_V^V	sums
C_K^1	n_{11}	n_{12}	$\dots$	n_{1V}	$a_1 = \sum_{i=1}^{V} n_{1i}$
C_K^2	n_{21}	n_{22}	$\dots$	n_{2V}	$a_2 = \sum_{i=1}^{V} n_{2i}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_K^K	n_{K1}	n_{K2}	$\dots$	n_{KV}	$a_K = \sum_{i=1}^{V} n_{Ki}$
sums	$b_1 = \sum_{j=1}^{K} n_{j1}$	$b_2 = \sum_{j=1}^{K} n_{j2}$	$\dots$	$b_V = \sum_{j=1}^{K} n_{jV}$	

The adjusted Rand Index between cluster structures $\boldsymbol{C}_K$ and $\boldsymbol{S}_V$ is computed as

$$\mathrm{ARI} = \frac{\sum\limits_{j=1}^{K}\sum\limits_{i=1}^{V}\binom{n_{ji}}{2} - \left[\sum\limits_{j=1}^{K}\binom{a_j}{2}\sum\limits_{i=1}^{V}\binom{b_i}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum\limits_{j=1}^{K}\binom{a_j}{2} + \sum\limits_{i=1}^{V}\binom{b_i}{2}\right] - \left[\sum\limits_{j=1}^{K}\binom{a_j}{2}\sum\limits_{i=1}^{V}\binom{b_i}{2}\right] \Big/ \binom{n}{2}}. \tag{2.22}$$

2.8 Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve was originally developed by radar engineers to detect enemy objects in battlefields (Collinson, 1998). A ROC curve is a graphical plot that evaluates the diagnostic ability of a binary classifier as its discrimination threshold is varied. For given input data which contain the correct (observed) labels for all observations, a binary classifier predicts outcomes with two class labels, e.g. 1/0 and Yes/No, where the classifier boundary between the two classes is determined by a threshold value. Commonly, the class of interest is denoted as “positive” and the other as “negative”.

A ROC curve is a plot of sensitivity against specificity at various discrimination threshold values. Here, sensitivity is calculated as the number of correct positive predictions divided by the total number of true positives, and specificity is calculated as the number of correct negative predictions divided by the total number of true negatives. Figure 2.2 shows an example of the ROC curve. The closer an ROC curve is to the upper left corner, the more accurate is the classifier. A common method to summarise the ROC performance is to compute the area under the curve, abbreviated AUC (Bradley, 1997), which is a single scalar value that measures the classifier performance across a number of possible threshold values. It takes values from 0 to 1, where an AUC of 1 corresponds to perfectly accurate classification, an AUC of 0.5 implies a random classification such that the ROC curve will fall on the diagonal (45-degree line), and an AUC of 0 indicates perfectly inaccurate classification.

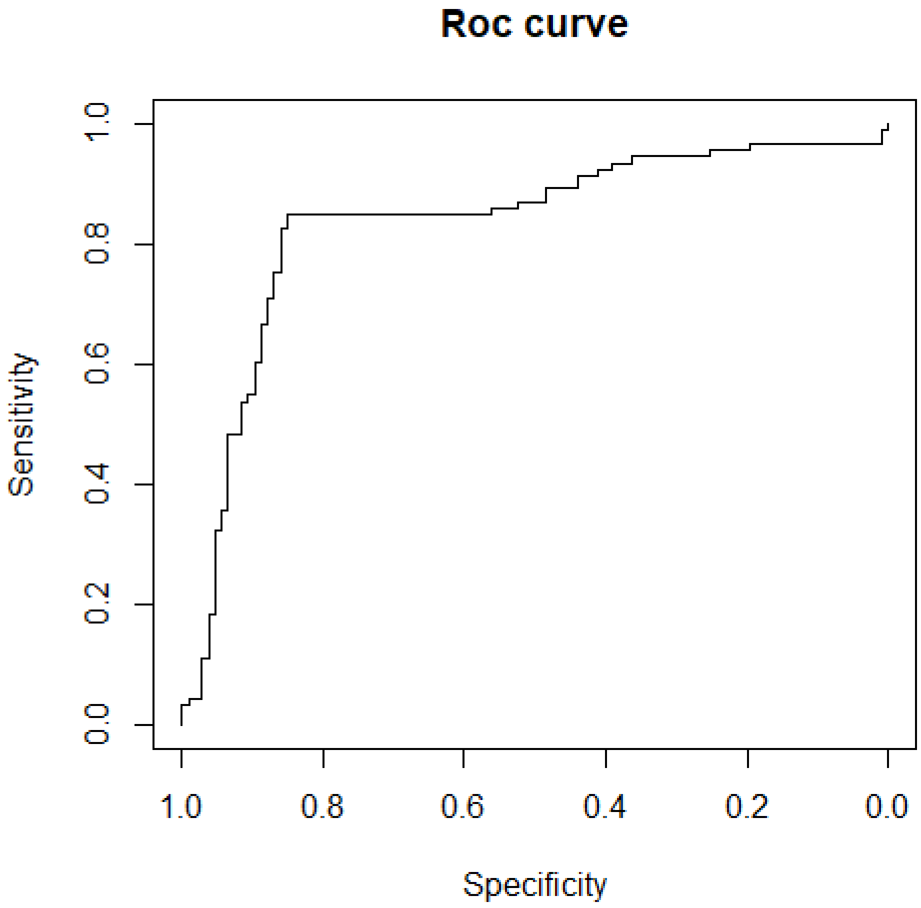


Figure 2.2: An example of the ROC curve.

Chapter 3

Estimating spatial disease risks and identifying clusters via a k-means clustering based approach

3.1 Introduction

Conditional autoregressive (CAR) models outlined in Section 2.4.4 are commonly used to capture the spatial autocorrelation present in areal unit count data when estimating the spatial pattern in disease risk. In these models spatial autocorrelation relating to n non-overlapping areal units that comprise the study region is induced by an $n \times n$ neighbourhood matrix $\boldsymbol{W} = \{w_{ij}\}$, which determines the degree of spatial closeness between pairs of areal units. As touched on in Section 2.4.2, the values of $\{w_{ij}\}$ are typically binary and determined by geographical adjacency in the literature, where $w_{ij} = 1$ if areal units (i, j) share a common border and $w_{ij} = 0$ otherwise (diagonal elements $w_{ii} = 0$ for all i). Using this border sharing rule data values relating to areas (i, j) will be modelled as being correlated as long as they border one another. Therefore such models assume a constant level of spatial autocorrelation across the entire study region, and hence a globally spatially smooth disease risk surface. However, this may not always be the case in practice because some pairs of geographically adjacent areas are likely to be independent of each other and exhibit significantly different disease risks due to factors such as population behaviours and socio-economic deprivation, for example, poor and rich areas which live side by side. Section 2.6.1 discussed a number of existing approaches that allow for spatial discontinuities in the disease risk surface, by identifying clusters of areas that exhibit elevated or reduced risks compared to their neighbours. Some of them force the clusters to be spatially contiguous, while some approaches do not. Additionally, some models assume constant disease risk within a cluster, while others are more flexible by allowing within cluster variation in disease risk.

In this chapter I propose new methodology to estimate the disease risk and identify spatial clusters of areas with high risks. The approach consists of two stages. In stage one the entire study region is partitioned into a set of clusters (or classes) of areal units in terms of their similarity in disease risk via a k-means clustering algorithm. The resultant cluster structure is used to decide whether or not the risks between geographically neighbouring areas should be smoothed over. This is achieved by changing the value of w_{ij} in the border sharing neighbourhood matrix $\boldsymbol{W}$ from 1 to 0 if areas (i, j) are geographically adjacent and in different clusters. With this, a set of candidate cluster structures as well as their corresponding candidate neighbourhood matrices are generated, with each reflecting a more realistic spatial pattern of disease risk via a different number of clusters. In stage two separate Bayesian hierarchical models are fitted for each candidate cluster structure/neighbourhood matrix, and the best model is then chosen using a model selection rule. I propose four potential approaches to select the best model which gives the most appropriate cluster structure/neighbourhood matrix from the candidates. The methodology is able to improve the estimation of the spatial pattern in disease risk, particularly when there are non-smooth disease risks present, e.g. risk discontinuities exist between neighbouring areas.

The remainder of this chapter is structured as follows. Section 3.2 presents our motivating data set for respiratory disease in Greater Glasgow in 2016. Section 3.3 outlines the proposed methodology, and its effectiveness is assessed against an existing and still widely used model by a large simulation study in Section 3.4. Section 3.5 applies the methodology to the motivating application, a study of respiratory disease risk in the 257 Intermediate Zones that comprise the Greater Glasgow and Clyde Health Board in 2016. Finally, the proposed methodology is further discussed in Section 3.6.

3.2 Motivating study

Respiratory disease is one of the leading causes of death in Scotland, and has International Classification of Disease tenth revision codes J00-J99 (https://www.nrscotland.gov.uk). The methodology is motivated by a study of respiratory disease risk in Glasgow, Scotland in 2016. As displayed in Figure 3.1, the study region is the Greater Glasgow and Clyde Health

Board which contains the city of Glasgow and the surrounding rural areas, and is split in two by the River Clyde running north west through the region. The study region has a population of around 1,200,000 people and is split up into $n = 257$ Intermediate Zones (IZs), each with an average population of approximately 4,000 residents. The disease data modelled here are available from Public Health Scotland. The response data, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, are the observed counts of the numbers of hospital admissions with a primary diagnosis of respiratory disease in 2016 for each of the 257 IZs, where Y_i represents the disease counts in IZ $i(= 1, \ldots, 257)$. The expected respiratory hospital admission counts for each IZ are calculated using indirect standardisation to adjust for different population sizes and age and sex structures across the IZs, and are denoted by $\boldsymbol{E} = (E_1, \ldots, E_n)$. These expected counts are based on Scotland-wide age-sex specific respiratory hospitalisation rates, because it allows us to examine how the risk of disease in Glasgow compares to the national average, which is the benchmark often used by Public Health Scotland when examining the spatial pattern in disease risk. In this study, the observed disease counts range between 21 and 256 in a single IZ with a median of 95, while the expected counts are between 16.59 and 135.55 in a single IZ with a median of 74.61.

The simplest measure of disease risk is the standardised incidence ratio (SIR), which is calculated as the ratio of the observed to the expected numbers of hospital admissions for each areal unit, i.e. $\text{SIR}_i = \frac{Y_i}{E_i}$. An SIR value greater than 1 corresponds to an increased level of disease risk within the areal unit compared to the Scottish average. In contrast a value less than 1 indicates a decreased level of risk compared to the average over Scotland. Figure 3.2 displays the SIR for respiratory disease hospitalisation in 2016 in Greater Glasgow. The median SIR over the 257 IZs is 1.28 with the maximum of 2.67 and minimum of 0.43. The figure shows that higher SIRs are mainly in the East End of Glasgow (the east of the map) and along the southern bank of the River Clyde. These high risk regions contain a number of socio-economically deprived areas such as Easterhouse, Govan, Barlanark and Paisley. However, the areas with low SIRs can be found in the center and far south of the city as well as in the outlying suburbs, e.g. Dowanhill, Eaglesham, Giffnock, Milngavie and Bearsden, which are the wealthy areas. These results suggest that the poor areas tend to exhibit higher SIRs than the affluent areas. In addition, there are numerous pairs of neighbouring areas where a discontinuity in disease risk appears to exist, suggesting the presence of clusters of areas that exhibit elevated risks compared with their neighbours. For example, in 2016 Drumchapel to the north west of the city exhibits a vastly higher SIR value (SIR = 2.59) compared to its neighbour Bearsden in the north east (SIR = 0.74). Therefore, the common approach in the literature of assuming that all pairs of neighbouring areal units are correlated and exhibit similar disease risks is clearly not appropriate, which motivates the spatial clustering model proposed in Section 3.3. The analysis of these motivating data aims at achieving a better estimation of the spatial pattern in respiratory disease risk in Greater Glasgow in 2016, and identifying the spatial extent of clusters of areas with elevated risks.

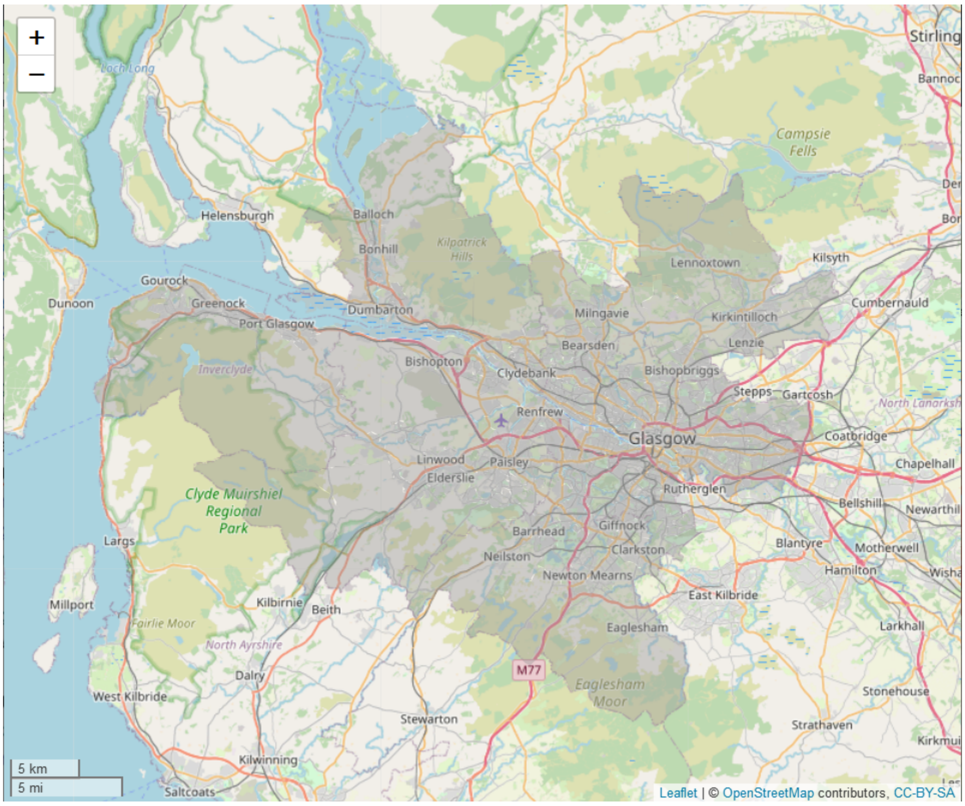


Figure 3.1: A map of the Greater Glasgow and Clyde Health Board (shaded region) overlaying on OpenStreetMap.

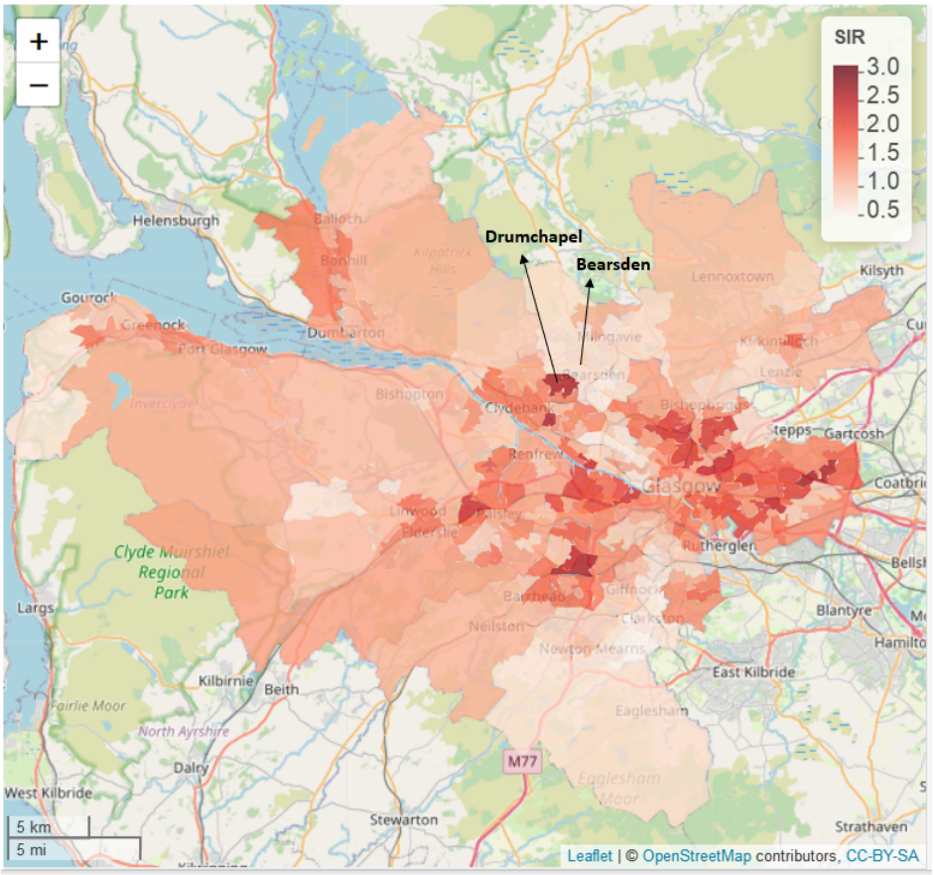


Figure 3.2: A map of the standardised incidence ratio (SIR) for respiratory disease in the Intermediate Zones in the Greater Glasgow and Clyde Health Board in 2016.

3.3 Methodology

I propose a two-stage modelling approach to estimate the spatial pattern in disease risk over the study region and detect clusters of areas with elevated or reduced disease risks. In the first stage all the areal units are split into k clusters using a k-means clustering algorithm, where k is an integer ranging from 1 to K. From this, K cluster structures are generated and further used to produce K candidate neighbourhood matrices accordingly. The optimal cluster structure and neighbourhood matrix of these candidates will be selected in the second stage through a model comparison procedure. In stage 2 a separate Bayesian hierarchical model is fitted using each candidate neighbourhood matrix that corresponds to a given cluster structure, and the best model will be selected via a model selection rule. Here, I propose four approaches, described in Section 3.3.3, to select the best model and their performances are assessed and compared in the simulation study in Section 3.4.

3.3.1 Notation

Consider a study region partitioned into n non-overlapping areal units indexed by $i \in \{1,\ldots,n\}$. $\boldsymbol{Y} = \{Y_i\}$ and $\boldsymbol{E} = \{E_i\}$ respectively denote the set of observed and expected disease counts in areal unit i, and a vector of covariates (if needed) is given by $\boldsymbol{x}_i$ for area i. This notation will be used in this chapter as well as in Chapter 4.

3.3.2 Stage 1 — Generating candidate neighbourhood matrices via k-means clustering

Given a value of k, i.e. a specified number of clusters in k-means clustering (MacQueen et al., 1967), the n non-overlapping areal units are partitioned into k clusters by clustering the natural log of the standardised incidence ratio (SIR) after adjusting for covariates, that is $\left\{\ln\left(\frac{Y_i}{E_i}\right) - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right\}$, where $\hat{\boldsymbol{\beta}}$ are initially estimated assuming independence via maximum likelihood estimation. Here, data are clustered on the logarithm scale of SIR because it corresponds to the linear predictor scale in model (3.1), and the covariate adjustment is because the clusters identified by our methodology are in the random effects surface in model (3.1)-(3.2). The resultant cluster structure is denoted by $\boldsymbol{C}_k = (C_k^1,\ldots,C_k^k)$, which indicates that the n areal units are divided into k clusters with different levels of average disease risk and C_k^j represents the j_{th} cluster in structure $\boldsymbol{C}_k$. Note that k-means clustering is applied to the data without regard to the spatial positions of the areal units, because the spatial correlation in the data is modelled by the spatial random effects in the proposed model (3.1). Thus the clusters identified here represent the number of different risk levels rather than the number of spatially distinct cluster. In addition, the clusters are ordered so that areal units in cluster C_k^j always have a lower average risk level than those in cluster C_k^{j+1} and a higher average risk level than those in cluster C_k^{j-1}. Each candidate cluster structure $\boldsymbol{C}_k$ is used to generate a candidate neighbourhood matrix denoted by $\boldsymbol{W}^{(k)}$. The k-means clustering algorithm and how it works to generate a candidate $\boldsymbol{W}^{(k)}$ are as follows:

K-means clustering algorithm

1. Specify the number of clusters k, for $k = 1,2,\ldots,K$ to consider. The value of K is chosen to be a sensible upper limit for the number of clusters one would expect to find in the data, which must be specified by the user. In this study I set $K = 10$ as a conservative overly large choice, because as described above this represents the number of distinct risk levels and not the number of spatially contiguous clusters.

2. Randomly select k data points from $(\xi_1, \ldots, \xi_n)$, where $\xi_i = \ln\left(\frac{Y_i}{E_i}\right) - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$, as initial cluster centers denoted by $(O_1, \ldots, O_k)$. The corresponding k clusters are denoted by $(C_k^1, \ldots, C_k^k)$.

3. Calculate the Euclidean distance between each areal unit and each cluster center, and then assign each areal unit to the cluster with the nearest distance. The Euclidean distance d_{ij} between areal unit i and cluster center O_j is computed as $d_{ij} = \sqrt{(\xi_i - O_j)^2}$.

4. Update the cluster centers $(O_1, \ldots, O_k)$ by taking the average of all areal units in each cluster, that is, $O_j = \frac{1}{n_j}\sum_{f:f \in C_k^j} \xi_f$, where n_j is the number of areal units in cluster C_k^j and $j = 1, \ldots, k$.

5. Iterate steps 3 and 4 until the cluster assignments do not change.

The cluster structure $\boldsymbol{C}_k = (C_k^1, \ldots, C_k^k)$ is used to create a candidate neighbourhood matrix $\boldsymbol{W}^{(k)}$, with each element given as

$$w_{ij}^{(k)} = \begin{cases} 1, & \text{if areal units } (i,\, j) \text{ share a common geographical border and are in the same cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

This algorithm leads to K candidate neighbourhood matrices in total, with each matrix corresponding to a cluster structure with a distinct number of clusters (or risk level). The estimation of $\left(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(K)}\right)$ in stage one is based on an Empirical Bayes approach, where we estimate the hyperparameter $\boldsymbol{W}$ from the data first, rather than assigning a hyperprior distribution to it (Robbins, 1992). In our case, we estimate $\boldsymbol{W}^{(k)}$ by clustering the data, and then apply a Bayesian hierarchical model described in Section 3.3.3 to the data and $\boldsymbol{W}^{(k)}$.

3.3.3 Stage 2 — Bayesian spatial modelling and model selection

Each value of k relates to a candidate neighbourhood matrix $\boldsymbol{W}^{(k)}$, therefore, we have K candidate neighbourhood matrices, $(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(K)})$. $\boldsymbol{W}^{(1)}$ is equal to the border sharing $\boldsymbol{W}$ as $k = 1$, which thus represents no clusters in disease risk. For each candidate neighbourhood matrix $\boldsymbol{W}^{(k)}$, a separate Bayesian hierarchical model is fitted to the data. The first level of the proposed spatial model is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i R_i^{(k)}), \ \ i = 1,\dots,n, \\ \ln(R_i^{(k)}) &= \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)} + \phi_i^{(k)}, \\ \beta_j^{(k)} &\sim \text{N}(0,1000), \text{ for } j = 0,\dots,p. \end{aligned} \tag{3.1}$$

Here $R_i^{(k)}$ is the disease risk in areal unit i relating to $\boldsymbol{W}^{(k)}$, and it can be estimated by two components. The first component is a $1 \times (p+1)$ vector of known covariates $\boldsymbol{x}_i^\top = (1, x_{i1}, \dots, x_{ip})$, including an intercept term, with regression parameters $\boldsymbol{\beta}^{(k)} = (\beta_0^{(k)}, \beta_1^{(k)}, \dots, \beta_p^{(k)})$. The prior specified for $\beta_j^{(k)}$ is a Gaussian prior distribution with mean zero and variance 1000. Note that the model outlined above is the most general form that includes covariates, but I do not include any covariate data in the respiratory disease motivating application in Section 3.5, because the aim of the analysis is to identify clusters in the disease risk surface, not in the residual surface after covariate adjustment. Furthermore, as the clusters identified by our methodology are in the random effects surface, then by not including covariates in the model the random effects surface and the disease risk surface have the same spatial structure, thus any clusters identified also relate to disease risk.

The second component is a set of spatial random effects $\boldsymbol{\phi}^{(k)} = (\phi_1^{(k)}, \dots, \phi_n^{(k)})$ that are used to account for the spatial autocorrelation in the data. These random effects can be modelled by a conditional autoregressive prior as discussed in Section 2.4.4. Here the random effects $\boldsymbol{\phi}^{(k)}$ are modelled by the Leroux CAR prior (Leroux et al., 2000), which is given by

$$\begin{aligned} \phi_i^{(k)} | \boldsymbol{\phi}_{-i}^{(k)} &\sim \text{N}\left(\frac{\rho^{(k)} \sum_{j=1}^n w_{ij}^{(k)} \phi_j^{(k)}}{\rho^{(k)} \sum_{j=1}^n w_{ij}^{(k)} + 1 - \rho^{(k)}}, \frac{\tau^{2^{(k)}}}{\rho^{(k)} \sum_{j=1}^n w_{ij}^{(k)} + 1 - \rho^{(k)}} \right), \\ \tau^{2^{(k)}} &\sim \text{Inverse-Gamma}(1, 0.01), \\ \rho^{(k)} &\sim \text{Uniform}(0,1), \end{aligned} \tag{3.2}$$

where $\boldsymbol{\phi}_{-i}^{(k)} = (\phi_1^{(k)}, \dots, \phi_{i-1}^{(k)}, \phi_{i+1}^{(k)}, \dots, \phi_n^{(k)})$. The parameter $\rho^{(k)}$ controls the level of spatial autocorrelation in the data, where $\rho^{(k)} = 0$ indicates independence in space (as $\phi_i^{(k)} \sim \text{N}(0, \tau^{2^{(k)}})$), while $\rho^{(k)} = 1$ indicates strong spatial dependence (corresponding to the intrinsic CAR prior (Besag et al., 1991)). $\tau^{2^{(k)}}$ is a variance parameter that controls the amount of spatial variation between the random effects. A weakly informative uniform prior on the interval [0,1] is assigned to $\rho^{(k)}$ and a conjugate Inverse-Gamma prior, Inverse-Gamma$(1,0.01)$, is assigned to $\tau^{2^{(k)}}$. To achieve identifiability, the spatial random effects are zero-mean centred.

The joint multivariate Gaussian distribution for $\boldsymbol{\phi}^{(k)}$ corresponding to the above Leroux prior is $\boldsymbol{\phi}^{(k)} \sim \mathrm{N}\left(\mathbf{0}, \tau^{2^{(k)}}\boldsymbol{Q}(\rho^{(k)}, \boldsymbol{W}^{(k)})^{-1}\right)$, where $\boldsymbol{Q}(\rho^{(k)}, \boldsymbol{W}^{(k)}) = \rho^{(k)}(\mathrm{diag}(\boldsymbol{W}^{(k)}\mathbf{1}) - \boldsymbol{W}^{(k)}) + (1-\rho^{(k)})\boldsymbol{I}$, $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{I}$ is an $n \times n$ identity matrix. The variance matrix between random effects $(\phi_i^{(k)}, \phi_j^{(k)})$ conditioning on the remaining random effects $\boldsymbol{\phi}_{-ij}^{(k)}$ is given by

$$\begin{aligned}
\mathrm{Var}\left[\phi_i^{(k)}, \phi_j^{(k)} | \boldsymbol{\phi}_{-ij}^{(k)}\right] &= \begin{bmatrix} \frac{\rho^{(k)}\sum_{v=1}^{n} w_{iv}^{(k)} + 1 - \rho^{(k)}}{\tau^{2^{(k)}}} & \frac{-\rho^{(k)} w_{ij}^{(k)}}{\tau^{2^{(k)}}} \\ \frac{-\rho^{(k)} w_{ji}^{(k)}}{\tau^{2^{(k)}}} & \frac{\rho^{(k)}\sum_{v=1}^{n} w_{jv}^{(k)} + 1 - \rho^{(k)}}{\tau^{2^{(k)}}} \end{bmatrix}_{2\times 2}^{-1} \\
&= \frac{\tau^{2^{(k)}}}{\Delta} \begin{bmatrix} \rho^{(k)}\sum_{v=1}^{n} w_{jv}^{(k)} + 1 - \rho^{(k)} & \rho^{(k)} w_{ij}^{(k)} \\ \rho^{(k)} w_{ji}^{(k)} & \rho^{(k)}\sum_{v=1}^{n} w_{iv}^{(k)} + 1 - \rho^{(k)} \end{bmatrix}_{2\times 2},
\end{aligned} \tag{3.3}$$

where Δ is computed as $\Delta = \left(\rho^{(k)}\sum_{v=1}^{n} w_{iv}^{(k)} + 1 - \rho^{(k)}\right)\left(\rho^{(k)}\sum_{v=1}^{n} w_{jv}^{(k)} + 1 - \rho^{(k)}\right) - \rho^{(k)^2} w_{ij}^{(k)} w_{ji}^{(k)}$. The partial correlation between $(\phi_i^{(k)}, \phi_j^{(k)})$ conditioning on the remaining effects $\boldsymbol{\phi}_{-ij}^{(k)}$, denoted by $\mathrm{Corr}\left(\phi_i^{(k)}, \phi_j^{(k)} | \boldsymbol{\phi}_{-ij}^{(k)}\right)$, can be derived as

$$\begin{aligned}
\mathrm{Corr}\left(\phi_i^{(k)}, \phi_j^{(k)} | \boldsymbol{\phi}_{-ij}^{(k)}\right) &= \frac{\mathrm{Cov}\left(\phi_i^{(k)}, \phi_j^{(k)} | \boldsymbol{\phi}_{-ij}^{(k)}\right)}{\sqrt{\mathrm{Var}\left(\phi_i^{(k)} | \boldsymbol{\phi}_{-i}^{(k)}\right) \mathrm{Var}\left(\phi_j^{(k)} | \boldsymbol{\phi}_{-j}^{(k)}\right)}} \\
&= \frac{\rho^{(k)} w_{ij}^{(k)}}{\sqrt{\left(\rho^{(k)}\sum_{v=1}^{n} w_{iv}^{(k)} + 1 - \rho^{(k)}\right)\left(\rho^{(k)}\sum_{v=1}^{n} w_{jv}^{(k)} + 1 - \rho^{(k)}\right)}}.
\end{aligned} \tag{3.4}$$

Equation (3.4) shows that $(\phi_i^{(k)}, \phi_j^{(k)})$ are only partially correlated if $w_{ij}^{(k)} = 1$, otherwise, the partial correlation between $(\phi_i^{(k)}, \phi_j^{(k)})$ is 0 and they are modelled as conditionally independent. Hence $\boldsymbol{W}^{(k)}$ determines the spatial correlation structure imposed by the model. The candidate neighbourhood matrices generated in stage one allow the random effects between neighbouring areas to be smoothed over only if they are in the same cluster, otherwise, the neighbouring areas are treated as conditionally independent and their values are not smoothed towards each other, which means that any cluster discontinuities in the spatial surface are not smoothed over in the estimation. Note that stage one could produce a candidate neighbourhood matrix where an areal unit has no neighbours due to it being a singleton cluster. The spatial random effects for these singletons are not allowed to smooth towards their geographically neighbours, because $\sum_{j=1}^{n} w_{ij}^{(k)} = 0$ for the area i in question and $\phi_i^{(k)} \sim \mathrm{N}(0, \frac{\tau^{2(k)}}{1-\rho^{(k)}})$.

The model described above is fitted to the data separately for each $\boldsymbol{W}^{(k)}$, with $k = 1, \ldots, K$. Then the K models are compared and the best model is selected using a model selection rule. The only changing variable across these models is the choice of $\boldsymbol{W}^{(k)}$ corresponding to a given cluster structure $\boldsymbol{C}_k$, therefore the model comparison procedure can be thought as a comparison of the candidate cluster structures. When we select the best model we are also selecting the most appropriate cluster structure, in other words, the most appropriate value of k. Four approaches are proposed to select the best model. The first three approaches work by comparing the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), which measures the relative fit of a set of Bayesian hierarchical models, and the last approach compares the effective number of independent parameters (p_d) of each candidate model, which is a measure of model complexity. Each of the four approaches is detailed below.

Suppose DIC_k and p_{d_k} respectively denote the DIC and effective number of parameters for the candidate model with the number of clusters k and neighbourhood matrix $\boldsymbol{W}^{(k)}$. The relative percentage difference in DIC from the model with k to $k+1$ clusters is denoted by diff_k and computed as $\mathrm{diff}_k = \frac{\mathrm{DIC}_k - \mathrm{DIC}_{k+1}}{\mathrm{DIC}_k} \times 100\%$. k^* is used to represent the most appropriate value of k, which implies the choice of the final model.

Approach I: k^* corresponds to the model with the minimum DIC, that is, $k^* = \underset{k}{\arg\min}(\mathrm{DIC}_k)$;

Approach II: k^* corresponds to the model with the maximum relative percentage of decrease in DIC, that is, $k^* = \underset{k}{\arg\max}(\mathrm{diff}_k)$+1;

Approach III: Define a threshold $c^* \in (0,1)$ for the relative percentage of decrease in DIC; if $\mathrm{diff}_{\underset{k}{\arg\max}(\mathrm{diff}_k)} \leq c^*$, then $k^* = \underset{k}{\arg\max}(\mathrm{diff}_k)$; if $\mathrm{diff}_{\underset{k}{\arg\max}(\mathrm{diff}_k)} > c^*$, we look up each of the values greater than $\underset{k}{\arg\max}(\mathrm{diff}_k)$, i.e. $\underset{k}{\arg\max}(\mathrm{diff}_k) + 1$, $\underset{k}{\arg\max}(\mathrm{diff}_k) + 2$, $\underset{k}{\arg\max}(\mathrm{diff}_k) + 3, \ldots, K$, until find the smallest value, say $k^{'}$, that have $\mathrm{diff}_{k^{'}} < c^*$, and then $k^* = k^{'}$. Different values of c^* are used in the study in

Section 3.4 to assess the influence of the threshold value on model performance, which are,

(a) $c^* = 5\%$;

(b) $c^* = 10\%$; and

(c) $c^* = 20\%$.

Approach IV: k^* corresponds to the model with the smallest p_d, that is, $k^* = \arg\min_k(p_{d_k})$.

The four approaches select the best model from different perspectives. **Approach I** selects the model with the cluster structure minimising the DIC. **Approach II** and **III** work by comparing the relative difference in DIC from the model with k to $k+1$ clusters, because we believe that as k reaches a good number of clusters for the data, increasing k further would not improve the estimation substantially and could even adversely affect the modelling performance due to deviation from the true data structure. **Approach IV** selects the model with the cluster structure minimising the p_d. This approach focuses on model complexity and aims to explain the data by using as few parameters as possible.

3.3.4 Inference

Model inference is performed in a Bayesian setting via Markov chain Monte Carlo (MCMC) simulation, using both the Metropolis-Hastings (Metropolis et al., 1953, Hastings, 1970) and Gibbs sampling steps (Geman and Geman, 1984). The MCMC algorithm is written and implemented in R (R Core Team, 2013). In order to speed up computation, the updates of random effects are written in C++ via the Rcpp package (Eddelbuettel et al., 2011, Eddelbuettel, 2013). In addition, as $\boldsymbol{W}^{(k)}$ is a sparse matrix, I exploit its triplet form to improve computational efficiency. Point estimates of the parameters are taken from the median of the posterior distribution of each model parameter. Convergence of the posterior samples is diagnosed by checking parameter trace plots and by Geweke diagnostics (Geweke, 1992).

3.4 Simulation Study

3.4.1 Aim

In this section, a simulation study is conducted to compare the performance of the model (3.1)-(3.2) proposed in the previous section when it is used in conjunction with different model selection rules for selecting the best cluster structure (see Section 3.3.3). Models P1, P2 and P4 denote the proposed model that uses **Approach I**, **II** and **IV** to choose the best cluster structure respectively, while models P3(a), P3(b) and P3(c) all use **Approach III** but each of them specifies a different threshold value, with $c^* = 5\%, 10\%$ and 20% respectively. These models are compared against an existing commonly used model in disease mapping, which is the Leroux CAR model (Leroux et al., 2000) outlined in Section 2.4.4. This model induces the spatial autocorrelation structure based on geographical adjacency via the border sharing $\boldsymbol{W}$.

3.4.2 Data generation

In order to make the simulation study as realistic as possible to the real data, disease data are simulated for the 257 Intermediate Zones (IZs) comprising the Greater Glasgow and Clyde Health Board. Clustered disease data are generated according to the template shown in Figure 3.3, which consists of three clusters of disease risk (high, medium and low risk levels) across the study region. The cluster structure template is chosen based on the SIR for respiratory disease admissions for IZs in Greater Glasgow in 2016 (see Figure 3.2). Specifically, areal units with the SIR above 0.7 quantile and below 0.3 quantile are assumed to have a high and low level of disease risk respectively and are shaded in red and blue, while the remaining areal units are in the medium-risk cluster and are shaded in grey.

Disease count data $\boldsymbol{Y} = \{Y_i\}$ are generated from the Poisson log-linear model (3.5) for the $n = 257$ IZs, and as previously described covariates are not included. The size of the expected disease counts $\boldsymbol{E} = \{E_i\}$ quantifies the disease prevalence and is varied to assess its influence on model performance. The expected disease counts $\boldsymbol{E}$ are uniformly drawn from three different intervals: $\boldsymbol{E} \in [10,30], [50,100]$, and $[100,150]$. The intercept term β_0 is fixed at 0, and the set of spatial random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are generated from a multivariate Gaussian distribution with a spatially correlated precision matrix proposed by Leroux et al. (2000) as $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$. Here $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{I}$ is an $n \times n$ identity matrix. $\boldsymbol{W}$ is the border sharing neighbourhood matrix corresponding to the study region and the spatial dependence parameter ρ is set equal to 0.99, which corresponds to strong spatial dependence. The three disease risk clusters are achieved by specifying a piecewise constant mean function $\boldsymbol{\mu} = \{-1, 0, 1\}$ for $\boldsymbol{\phi}$ following the template shown in Figure 3.3. Values $\boldsymbol{\mu} = \{-1, 0, 1\}$ are multiplied by a constant scalar Z to adjust the magnitude of the differences between clusters, where larger values represent larger differences in disease risk. Values of $Z = 1, 0.5, 0$ are used in this study, where $Z = 1, 0.5$ respectively correspond to large and small differences between the clusters in the spatial surface, and $Z = 0$ corresponds to a spatially smooth risk surface with no clusters so that all areal units in the study region have the same expectation of disease risk. Therefore, the simulation study is split into nine different scenarios comprising pairwise combinations of $Z = 1, 0.5, 0$ and $\boldsymbol{E} \in [10, 30], [50, 100]$ and $[100, 150]$. The data are generated by

$$\begin{aligned} Y_i | E_i, R_i &\sim \text{Poisson}(E_i R_i), \quad i = 1, \ldots, n, \\ \ln(R_i) &= \beta_0 + \phi_i, \\ \boldsymbol{\phi} &\sim \text{N}(\boldsymbol{\mu}, \tau^2 \boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}). \end{aligned} \tag{3.5}$$

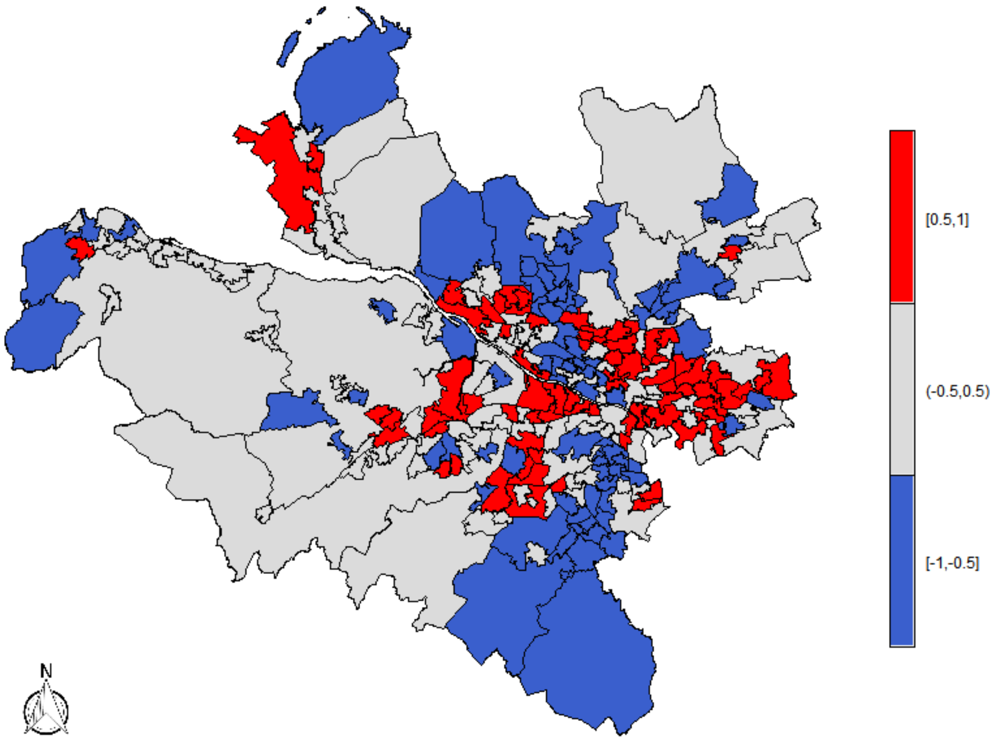


Figure 3.3: A map of the simulated cluster structure in the Greater Glasgow and Clyde Health Board. High-risk, medium-risk and low-risk clusters are respectively shaded in red, grey and blue.

3.4.3 Results

Two hundred simulated data sets are generated under each of the nine scenarios, and seven models, containing models P1, P2, P3(a), P3(b), P3(c), P4 and the Leroux model, are fitted to each data set. In all scenarios inference for each model is based on 100,000 MCMC samples with a burn-in period of 80,000. The Markov chain is thinned by 10 due to limited computer memory capacity and to reduce autocorrelation, which yields a total of 2,000 posterior samples.

The relative performances of the fitted models are compared using five metrics, and the results of the study for all nine scenarios are respectively summarised in Figures 3.4, 3.5 and 3.6, and outlined in Tables 3.1 and 3.2. The accuracy of disease risk estimation is quantified by the bias, root mean square error (RMSE) and coverage probabilities of the 95% credible intervals for the corresponding risk estimates, which are outlined as follows.

Bias

Bias measures the average difference between the estimated and the true values. The bias of the risk estimates of all areal units for each data set is calculated as

$$\text{Bias}(R_1,\ldots,R_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{R}_i - R_i), \tag{3.6}$$

where $\hat{R}_i$ represents the estimate of the true risk R_i for areal unit i.

Root mean square error (RMSE)

RMSE quantifies the average magnitude of the differences between the estimated and the true values. A lower RMSE value suggests a more accurate estimation. The RMSE of the risk estimates of all areal units for each data set is calculated as

$$\text{RMSE}(R_1,\ldots,R_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{R}_i - R_i)^2}. \tag{3.7}$$

Coverage probability

The uncertainty of the estimates can be measured by the coverage probabilities of the 95% credible intervals. The 95% coverage probability of risk R_i is computed as the proportion of the 95% credible intervals for R_i that contain the true value for R_i.

The correctness of the estimated cluster structures of the proposed models is measured by both the estimated number of clusters and the adjusted Rand Index between the true and estimated cluster structures. The adjusted Rand Index (ARI) proposed by Hubert and Arabie (1985) is a measure of the similarity between two cluster structures. A value of 1 indicates complete agreement between two cluster structures, a value of 0 indicates that the data points are randomly allocated to the two cluster structures, and a value less than 0 indicates that the level of agreement between the two cluster structures is smaller than that arising from randomly allocated data points. More information on the ARI can be found in Section 2.7.3.

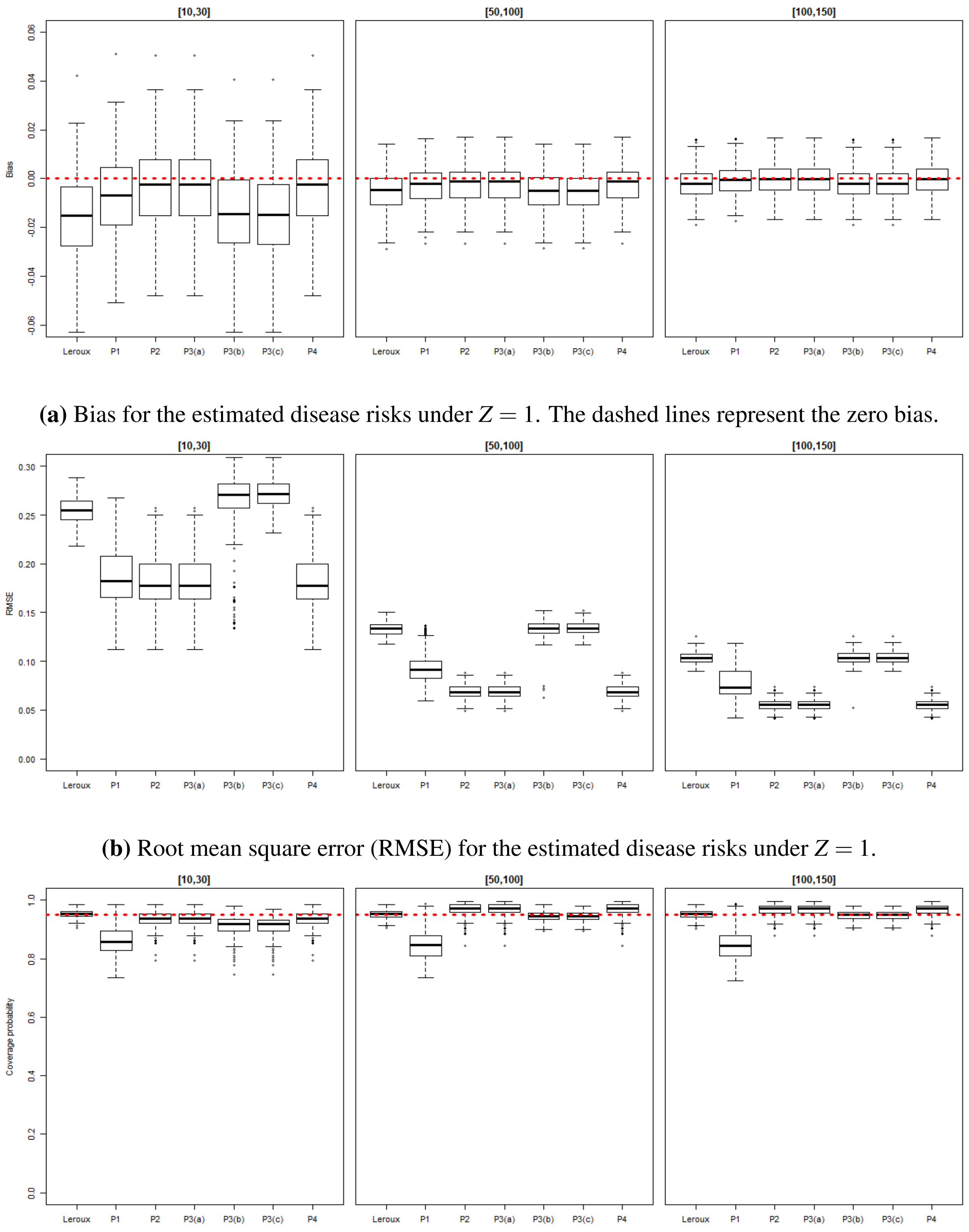
Figures 3.4, 3.5 and 3.6 display boxplots of the performance metrics by each model under different scenarios of $Z = 1, 0.5, 0$ over all simulated data sets. The results indicate that **Approach III** is not a sensible model selection rule, because the model performance is not robust to the choice of threshold c^* for all scenarios. Model P3(a) (where $c^* = 5\%$) outperforms P3(b) (where $c^* = 10\%$) and P3(c) (where $c^* = 20\%$) in terms of risk estimation when $Z = 1, 0.5$, whereas it exhibits higher RMSE values than the latter two when $Z = 0$, suggesting that the accuracy of risk estimates is affected by the threshold value which is unknown.

Figure 3.4 shows that when $Z = 1$ the proposed models P1, P2 and P4 perform better than the Leroux model under all three cases of $\boldsymbol{E}$ in terms of lower RMSE values and negligible bias close to zero; under $\boldsymbol{E} \in [10, 30]$ the median RMSE values are 0.183, 0.178, 0.178 for P1, P2 and P4, and 0.255 for the Leroux model, under $\boldsymbol{E} \in [50, 100]$ the median RMSE values are 0.091, 0.068, 0.068 for P1, P2 and P4, and 0.134 for the Leroux, and under $\boldsymbol{E} \in [100, 150]$ the median RMSE values are 0.073, 0.055, 0.055 for P1, P2 and P4, and 0.103 for the Leroux. Models P2, P4 and the Leroux all exhibit good coverage ability since the 95% credible intervals are able to contain the true risks around 95% of the time, while model P1 gives a slightly lower coverage probability around 0.85. It should also be noted that model P1 slightly overestimates the number of clusters with a median of 5 clusters, which unsurprisingly results in the relatively lower ARI values compared to models P2 and P4.

For the scenario of $Z = 0.5$, a more difficult case with smaller differences between clusters, Figure 3.5 shows that models P1, P2 and P4 outperform the Leroux model under larger values of $\boldsymbol{E}$, with the reductions in the median RMSE being 12.07%, 24.14%, 24.14% when $\boldsymbol{E} \in [50,100]$ and 16.48%, 37.36%, 37.36% when $\boldsymbol{E} \in [100,150]$ respectively. In addition, models P2 and P4 can accurately identify clusters and obtain higher ARI values close to 1 than model P1. However, when $\boldsymbol{E} \in [10,30]$, our models have slightly higher RMSE values, with medians of 0.216, 0.235 and 0.230 for P1, P2 and P4 respectively compared to 0.202 for the Leroux model, which is likely to be a result of poor clustering performance (small ARI values).

Table 3.2 shows that when there are clusters present in the data $(Z = 1, 0.5)$, models P1, P2 and P4 generally perform better if the expected disease counts are higher. This is because $\boldsymbol{Y} \propto \boldsymbol{E}\exp(\boldsymbol{\phi})$, and multiplying the fixed $\boldsymbol{\phi}$ with small values of $\boldsymbol{E}$ (e.g. $\boldsymbol{E} \in [10,30]$) would make the difference in risk between areas much less prominent in terms of the size of the difference in $\boldsymbol{Y}$. As a result, disease clusters are hard to accurately identify for rare diseases with small values of $\boldsymbol{E}$, and any incorrect cluster structure could make the smooth of random effects between neighbours unreliable and influence the veracity of the modelling results.

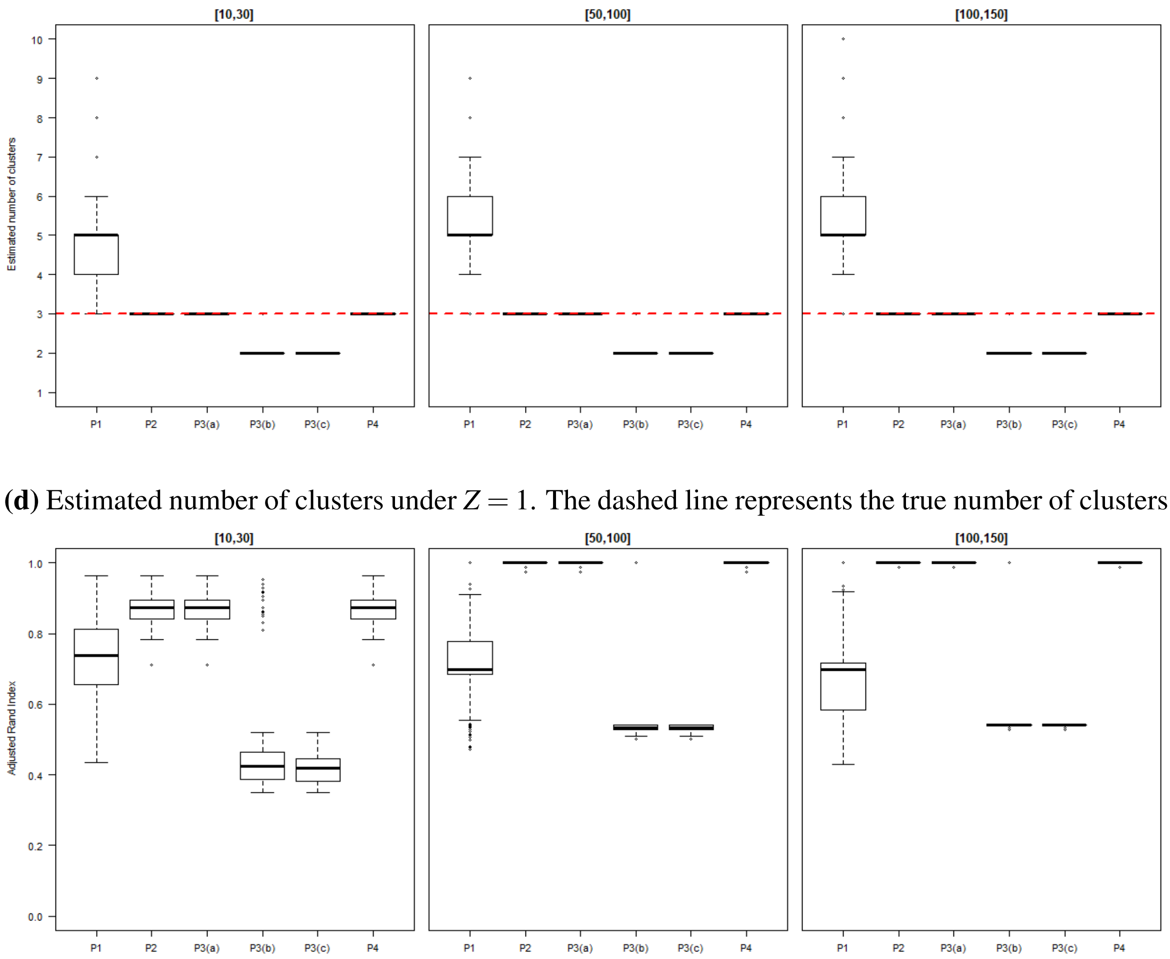
When the simulated risk surface is globally spatially smooth with no clusters, that is $Z = 0$, Table 3.1 and Figure 3.6 show that model P4 is able to estimate the correct cluster structure with ARI values equal to 1, and performs as good as the Leroux model under all three cases of $\boldsymbol{E}$ in terms of RMSE, bias and coverage probability. However, models P1 and P2 overestimate the number of clusters and provide slightly higher RMSE values and poorer coverage probabilities compared with P4 and the Leroux. In summary, the proposed model P4 performs consistently well, in particular outperforming the Leroux model. The only slight exception to this is when the disease is rare ($\boldsymbol{E} \in [10,30]$) and the clusters are not large in size ($Z = 0.5$), which is the case where the clusters are hardest to identify and hence all the cluster models perform less well. Models P1 and P2 provide less accurate risk estimates and cluster structures when $Z = 0$ than model P4, therefore they are only suitable for the case when there are clusters present in the data, but model P2 is preferable to P1 because it has smaller RMSE values and lower variability in the ARI values. Model P3 is not recommended in any circumstances because its performance has been shown to be dependent on the unknown threshold value.



(a) Bias for the estimated disease risks under $Z = 1$. The dashed lines represent the zero bias.

(b) Root mean square error (RMSE) for the estimated disease risks under $Z = 1$.

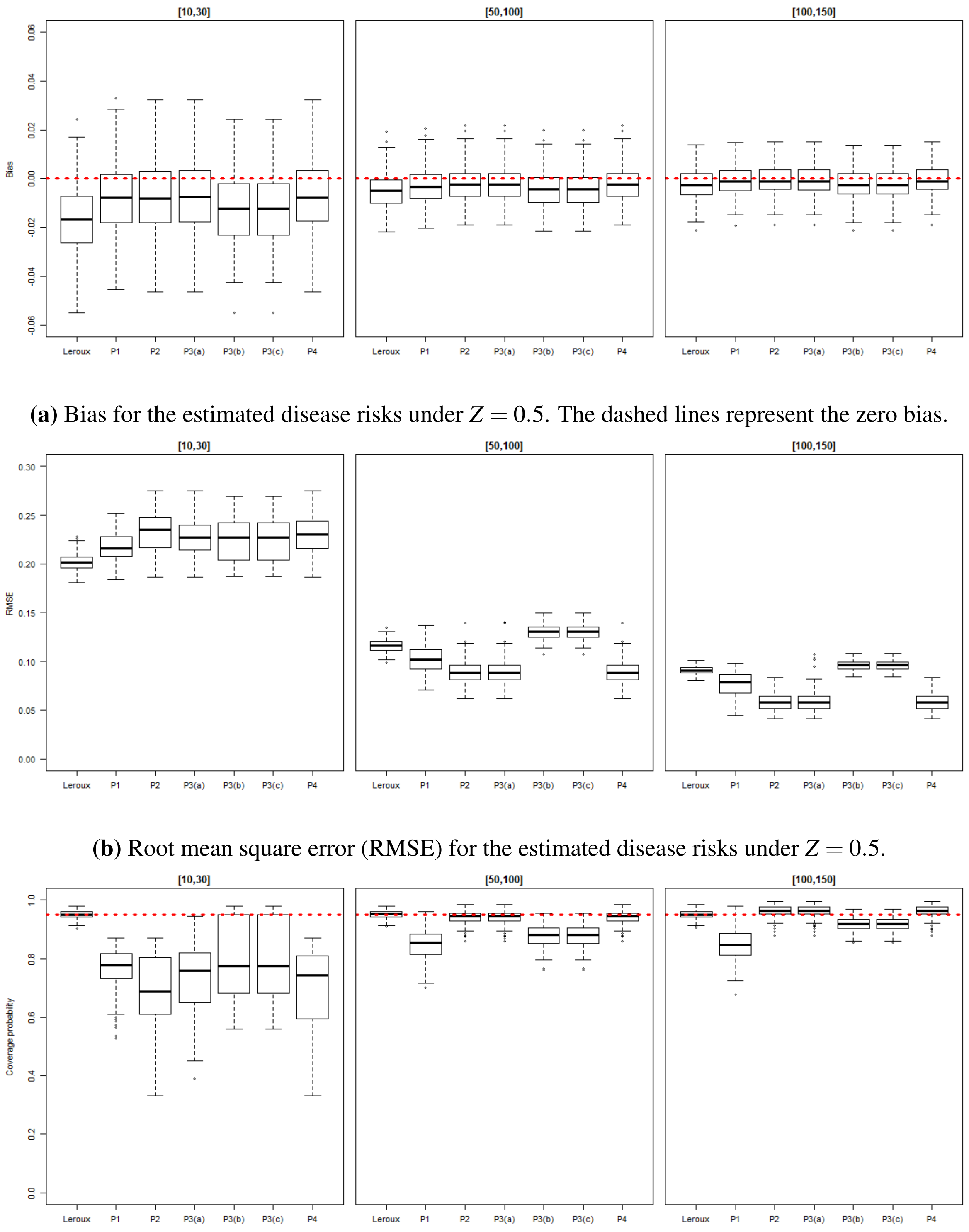
(c) Coverage probability of the 95% credible intervals for the estimated disease risks under $Z = 1$. The dashed lines represent the nominal 0.95 coverage levels.



(d) Estimated number of clusters under $Z = 1$. The dashed line represents the true number of clusters.

(e) Adjusted Rand Index (ARI) under $Z = 1$.

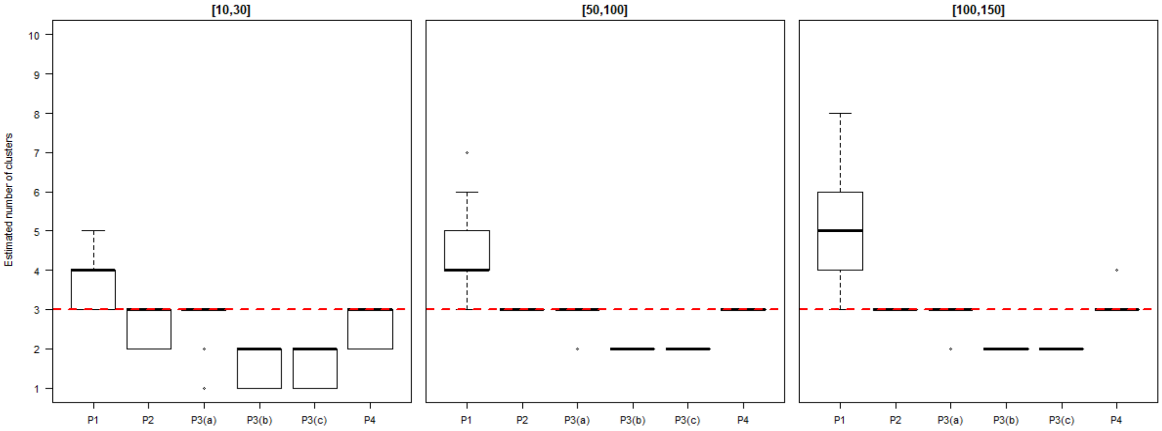
Figure 3.4: Simulation study results from $\boldsymbol{E} \in [10,30]$, $[50,100]$ and $[100,150]$ in terms of bias (3.4(a)), RMSE (3.4(b)), 95% coverage probabilities (3.4(c)), the estimated number of clusters (3.4(d)) and the adjusted Rand Index (3.4(e)) for each model under the scenario $Z = 1$.



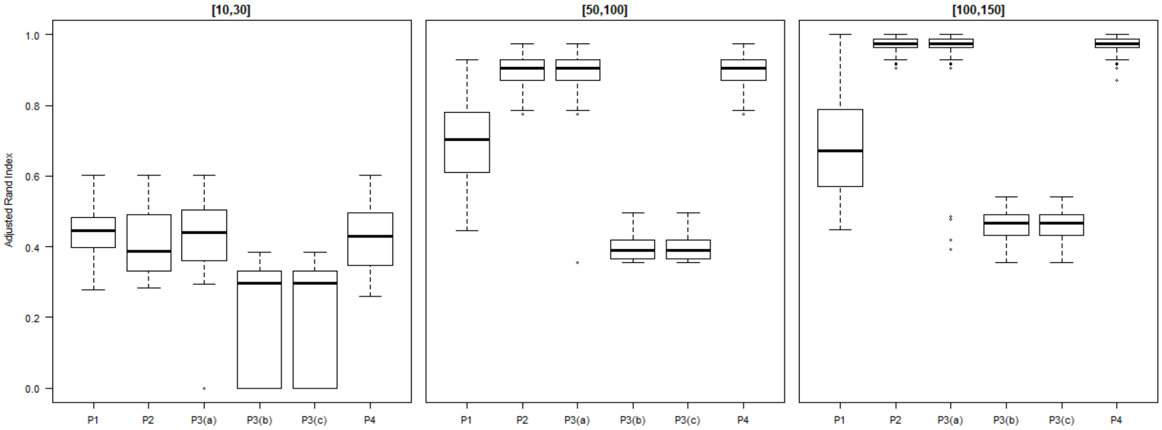
(a) Bias for the estimated disease risks under $Z = 0.5$. The dashed lines represent the zero bias.

(b) Root mean square error (RMSE) for the estimated disease risks under $Z = 0.5$.

(c) Coverage probability of the 95% credible intervals for the estimated disease risks under $Z = 0.5$. The dashed lines represent the nominal 0.95 coverage levels.

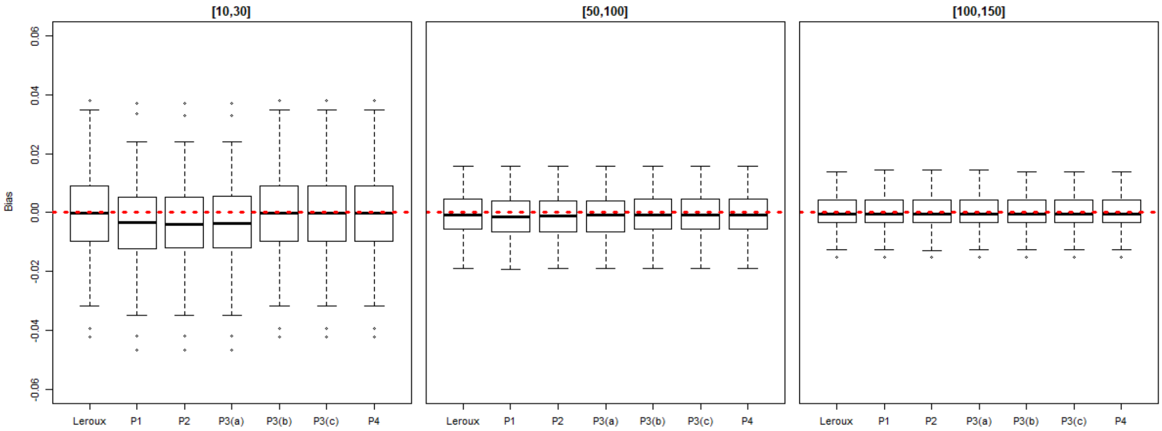


(d) Estimated number of clusters under $Z = 0.5$. The dashed line represents the true number of clusters.

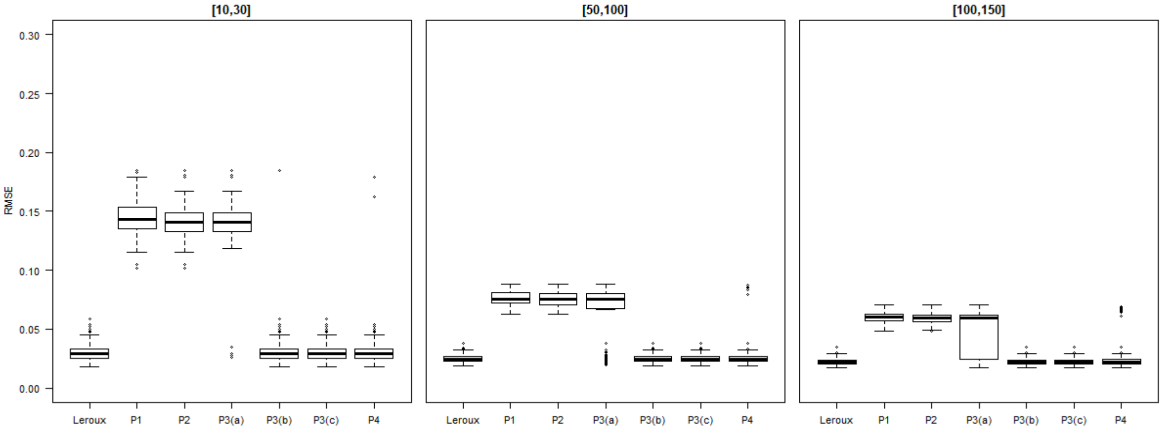


(e) Adjusted Rand Index (ARI) under $Z = 0.5$.

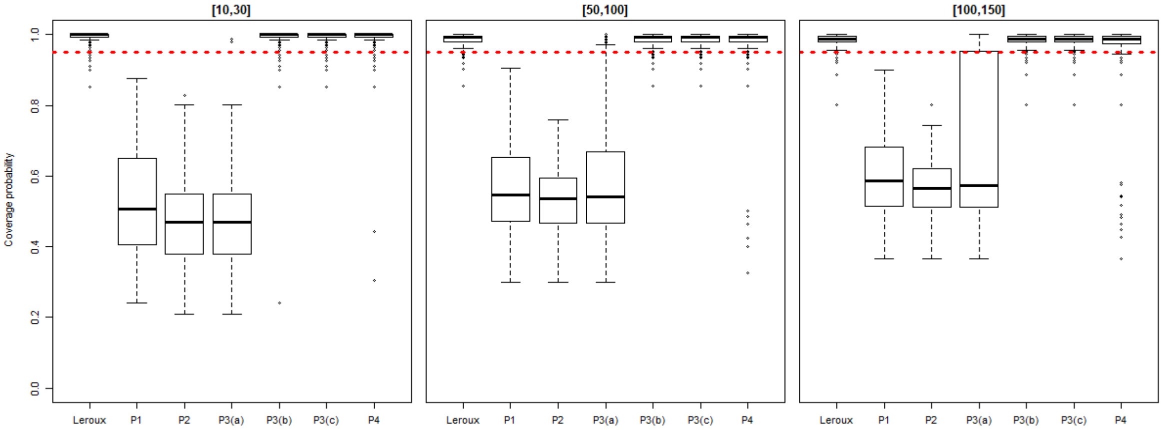
Figure 3.5: Simulation study results from $\boldsymbol{E} \in [10,30]$, $[50,100]$ and $[100,150]$ in terms of bias (3.5(a)), RMSE (3.5(b)), 95% coverage probabilities (3.5(c)), the estimated number of clusters (3.5(d)) and the adjusted Rand Index (3.5(e)) for each model under the scenario $Z = 0.5$.



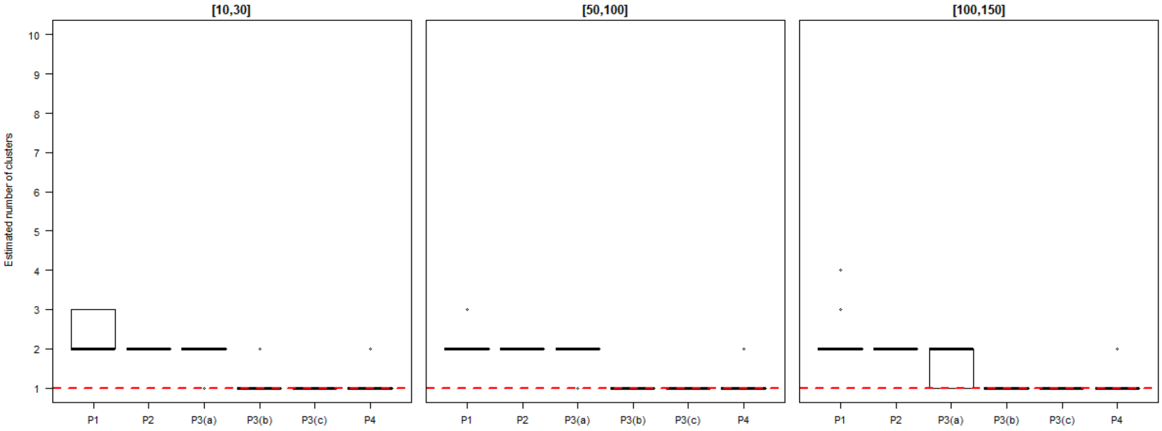
(a) Bias for the estimated disease risks under $Z = 0$. The dashed lines represent the zero bias.



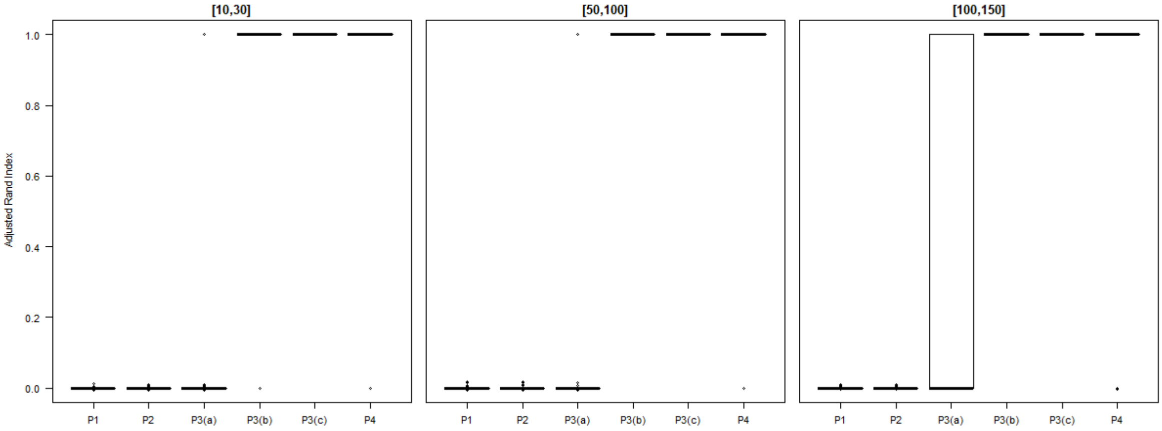
(b) Root mean square error (RMSE) for the estimated disease risks under $Z = 0$.



(c) Coverage probability of the 95% credible intervals for the estimated disease risks under $Z = 0$. The dashed lines represent the nominal 0.95 coverage levels.



(**d**) Estimated number of clusters under $Z = 0$. The dashed line represents the true number of clusters.



(**e**) Adjusted Rand Index (ARI) under $Z = 0$.

Figure 3.6: Simulation study results from $\boldsymbol{E} \in [10,30]$, $[50,100]$ and $[100,150]$ in terms of bias (3.6(a)), RMSE (3.6(b)), 95% coverage probabilities (3.6(c)), the estimated number of clusters (3.6(d)) and the adjusted Rand Index (3.6(e)) for each model under the scenario $Z = 0$.

Table 3.1: Summary of the median bias, RMSE and coverage probability of the 95% credible intervals of the estimated risk surface for each model. Values in brackets display the standard deviation. Note that for each of the performance metrics, the best results under each scenario of $\boldsymbol{E}$ and Z are highlighted in bold.

			Model						
Metric	$\boldsymbol{E}$	Z	Leroux	P1	P2	P3(a)	P3(b)	P3(c)	P4
Bias	[10, 30]	1	-0.015 (0.017)	-0.007 (0.017)	**-0.002 (0.017)**	**-0.002 (0.017)**	-0.015 (0.018)	-0.015 (0.017)	**-0.002 (0.016)**
		0.5	-0.017 (0.015)	-0.008 (0.015)	-0.008 (0.015)	**-0.007 (0.015)**	-0.012 (0.015)	-0.012 (0.015)	-0.008 (0.015)
		0	**0.000 (0.013)**	-0.003 (0.014)	-0.004 (0.014)	-0.004 (0.014)	**0.000 (0.013)**	**0.000 (0.013)**	**0.000 (0.013)**
	[50, 100]	1	-0.005 (0.008)	-0.002 (0.008)	**-0.001 (0.008)**	**-0.001 (0.008)**	-0.005 (0.008)	-0.005 (0.008)	**-0.001 (0.008)**
		0.5	-0.005 (0.008)	-0.003 (0.008)	**-0.002 (0.008)**	**-0.002 (0.008)**	-0.004 (0.008)	-0.004 (0.008)	**-0.002 (0.008)**
		0	**-0.001 (0.007)**	**-0.001 (0.007)**	**-0.001 (0.007)**	**-0.001 (0.007)**	**-0.001 (0.007)**	**-0.001 (0.007)**	**-0.001 (0.007)**
	[100, 150]	1	-0.002 (0.006)	**0.000 (0.006)**	**0.000 (0.006)**	**0.000 (0.006)**	-0.002 (0.006)	-0.002 (0.006)	**0.000 (0.006)**
		0.5	-0.003 (0.006)	**0.001 (0.006)**	**0.001 (0.006)**	**0.001 (0.006)**	-0.003 (0.006)	-0.003 (0.006)	**-0.001 (0.006)**
		0	**0.000 (0.005)**	**0.000 (0.005)**	**0.000 (0.005)**	**0.000 (0.005)**	**0.000 (0.005)**	**0.000 (0.005)**	**0.000 (0.005)**
RMSE	[10, 30]	1	0.255 (0.014)	0.183 (0.031)	**0.178 (0.027)**	**0.178(0.027)**	0.271 (0.040)	0.271 (0.016)	**0.178 (0.027)**
		0.5	**0.202 (0.009)**	0.216 (0.014)	0.235 (0.019)	0.227 (0.019)	0.227 (0.022)	0.227 (0.022)	0.230 (0.018)
		0	**0.029 (0.007)**	0.143 (0.013)	0.141 (0.012)	0.141 (0.020)	**0.029 (0.013)**	**0.029 (0.008)**	**0.029 (0.016)**
	[50, 100]	1	0.134 (0.007)	0.091 (0.019)	**0.068 (0.007)**	**0.068 (0.007)**	0.134 (0.012)	0.134 (0.007)	**0.068 (0.007)**
		0.5	0.116 (0.006)	0.102(0.014)	**0.088 (0.013)**	**0.088 (0.013)**	0.131 (0.008)	0.131 (0.008)	**0.088 (0.013)**
		0	**0.024 (0.003)**	0.076 (0.006)	0.075 (0.006)	0.075 (0.023)	**0.024 (0.003)**	**0.024 (0.003)**	**0.024 (0.011)**
	[100, 150]	1	0.103 (0.006)	0.073 (0.016)	**0.055 (0.005)**	**0.055 (0.005)**	0.104 (0.007)	0.104 (0.006)	**0.055 (0.005)**
		0.5	0.091(0.004)	0.079 (0.012)	**0.057 (0.009)**	0.058 (0.010)	0.096 (0.005)	0.096 (0.005)	**0.057 (0.009)**
		0	**0.022 (0.003)**	0.060 (0.004)	0.060 (0.004)	0.060 (0.018)	**0.022 (0.003)**	**0.022 (0.003)**	**0.022 (0.011)**
Coverage probability	[10, 30]	1	**0.953**	0.858	0.938	0.938	0.918	0.918	0.938
		0.5	**0.949**	0.778	0.687	0.759	0.774	0.774	0.743
		0	**1**	0.506	0.469	0.469	**1**	**1**	**1**
	[50, 100]	1	0.953	0.848	**0.973**	**0.973**	0.946	0.946	**0.973**
		0.5	**0.953**	0.856	0.946	0.946	0.881	0.881	0.946
		0	**0.992**	0.547	0.537	0.541	**0.992**	**0.992**	**0.992**
	[100, 150]	1	0.953	0.844	**0.973**	**0.973**	0.949	0.949	**0.973**
		0.5	0.949	0.848	**0.965**	**0.965**	0.918	0.918	**0.965**
		0	**0.988**	0.586	0.564	0.574	**0.988**	**0.988**	**0.988**

Table 3.2: Summary of the median estimated number of clusters and adjusted Rand Index (ARI) for each proposed model. The true number of clusters is 3 for $Z = 1$ and $Z = 0.5$, and 1 for $Z = 0$. Note that the best results for each metric under each scenario of $\boldsymbol{E}$ and Z are highlighted in bold.

			Model					
Metric	$\boldsymbol{E}$	Z	P1	P2	P3(a)	P3(b)	P3(c)	P4
Estimated number of clusters	[10, 30]	1	5	**3**	**3**	2	2	**3**
		0.5	4	**3**	**3**	2	2	**3**
		0	2	2	2	**1**	**1**	**1**
	[50, 100]	1	5	**3**	**3**	2	2	**3**
		0.5	4	**3**	**3**	2	2	**3**
		0	2	2	2	**1**	**1**	**1**
	[100, 150]	1	5	**3**	**3**	2	2	**3**
		0.5	5	**3**	**3**	2	2	**3**
		0	2	2	2	**1**	**1**	**1**
Adjusted Rand Index	[10, 30]	1	0.737	**0.873**	**0.873**	0.423	0.419	**0.873**
		0.5	**0.446**	0.385	0.441	0.297	0.297	0.430
		0	0.000	0.000	0.000	**1.000**	**1.000**	**1.000**
	[50, 100]	1	0.699	**1.000**	**1.000**	0.534	0.534	**1.000**
		0.5	0.702	**0.906**	**0.906**	0.389	0.389	**0.906**
		0	0.000	0.000	0.000	**1.000**	**1.000**	**1.000**
	[100, 150]	1	0.697	**1.000**	**1.000**	0.541	0.541	**1.000**
		0.5	0.670	**0.976**	**0.976**	0.467	0.467	**0.976**
		0	0.000	0.000	0.000	**1.000**	**1.000**	**1.000**

3.5 Application to real data

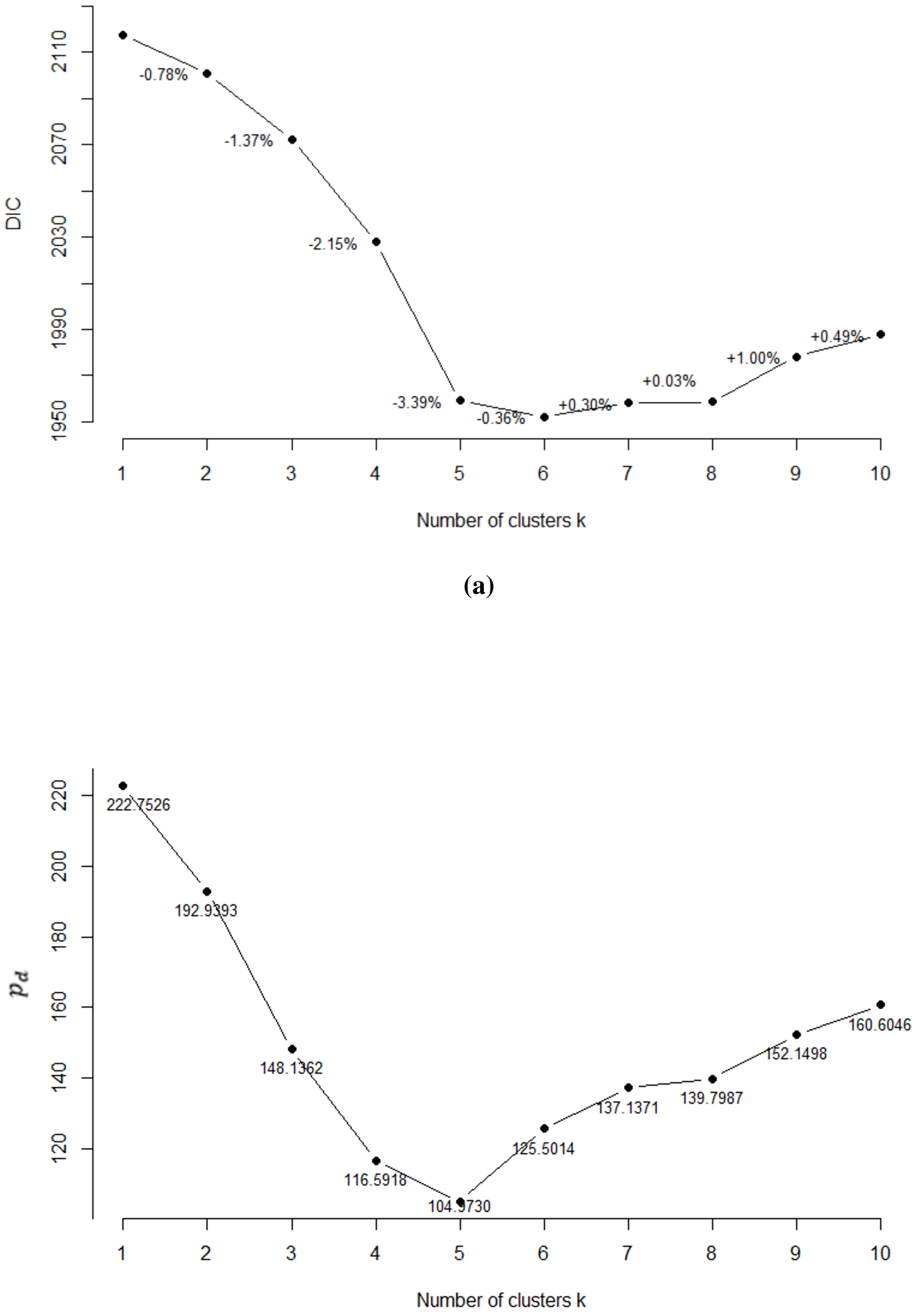
The simulation study has shown that model P4 (i.e. corresponding to **Approach IV** which selects the cluster structure minimising the p_d) appears to be the best fitting model, therefore here I only present the results of applying model P4 to the motivation study described in Section 3.2, which is a study of the respiratory disease risk in Greater Glasgow in 2016. The study region is the Greater Glasgow and Clyde Health Board (see Figure 3.1) and the respiratory disease data were introduced in Section 3.2. Posterior inference is based on a Markov chain with 100,000 samples, 80,000 of which were discarded for the burn-in period and the remaining 20,000 samples were thinned by 10.

Figure 3.7 displays both the DIC and p_d values for modelling the data by model (3.1)-(3.2) with the number of clusters k varying from 1 to $K = 10$. By design, model P4 identifies the most appropriate cluster structure as the one having 5 distinct clusters, because it has the minimum p_d which is 104.973. As previously stated, by clusters we mean the number of non-spatial clusters (distinct risk levels) and not the number of spatially contiguous clusters. The estimated cluster structure also produces a good model fit to the data due to the relatively low DIC value. Note that although the structure with 6 clusters has a marginally lower DIC than the estimated structure, it models the data using more effective parameters.

Figure 3.8 displays the map of the five clusters of areas with different risk levels in the estimated cluster structure in Greater Glasgow, although from the map it is clear that there are many more spatially distinct clusters. Areas with darker shading exhibit a higher level of disease risk and the spatial discontinuities in the risk surface exist between geographically adjacent areas that are in different clusters. Clusters 4 and 5 have moderately high and high disease risks, with mean values of 1.66 and 2.24 respectively. The high-risk clusters are mainly in the east of Glasgow and along the southern bank of the River Clyde as well as some areas to the south of the river. In contrast, low-risk clusters (clusters 1 and 2) mostly lie in the center of northern Glasgow and in the outlying rural areas, with a mean risk of 0.69 for cluster 1 and 0.97 for cluster 2.

Figure 3.9 maps the respiratory disease risk estimates in Greater Glasgow in 2016 from model P4, where the arrows in the map identify some typical areas mentioned in this section. The estimated spatial risk pattern is similar to the SIR map displayed in Figure 3.2.

In 2016 the mean risk across Greater Glasgow is 1.342, suggesting that on average the respiratory disease risk in Greater Glasgow is about 34.2% higher than the Scottish average. The risk surface is not spatially smooth in that some areas exhibit notably different risks from their neighbours, which suggests the presence of clusters. The four areas with the highest risk estimates are highlighted in the map, including Yoker, Drumry, Drumchapel and Nitshill whose risks are 2.56, 2.53, 2.53 and 2.50 respectively. Dowanhill, an upper middle-class affluent residential district, has the risk as low as 0.652. One of the inducing factors that explain this spatial variation in risk is socio-economic deprivation, which has been widely justified to have a non-negligible influence on ill health (McCartney, 2012). The areas with higher risks are predominantly located in the East End of the city (which are roughly surrounded by a rectangle in the risk map) such as Springburn, Easterhouse and Barlanark, and along the southern bank of the Clyde river such as Govan area, which all typically exhibit high levels of socio-economic deprivation. The lower risk areas are wealthier and mostly in the affluent West End of Glasgow (just the north of the river) e.g. Hillhead and Dowanhill, and in the south of the city centre e.g. Newton Mearns and Clarkston. Rural areas with small populations and good living environment also tend to have low risks, for example Milngavie, which is a popular retirement place at the northwestern edge of Glasgow, and Eaglesham to the extreme south east. Figure 3.10 displays the estimated spatial risk pattern from the Leroux CAR model. The risk estimates from the proposed model and the Leroux model are similar, with a mean absolute difference in the posterior median risk estimates of 0.069. In addition, the Leroux model appears to induce increased levels of smoothing between those neighbouring areas that have been identified as being in different clusters by model P4. For example, Figure 3.10 shows that the color shades in the adjacent areas Kilmacolm and Renfrewshire Rural North are more similar to each other than those in Figure 3.9, with a risk estimate difference between the two areas of 0.04 for the Leroux model compared to 0.26 for model P4. Other examples include neighbouring areas Lennoxtown and Torrance, with a risk estimate difference of 0.03 for the Leroux model compared to 0.30 for model P4, as well as areas Kirkintilloch and Rosebank, with a risk difference of 0.62 for the Leroux model compared to 0.85 for model P4. This phenomenon is to be expected because in the Leroux model smoothing is with all neighbouring areas, while in my model the spatial random effects are not allowed to smooth towards their geographically neighbours that are in different clusters.



(a)

(b)

Figure 3.7: Plots of the Deviance Information Criterion (DIC) (3.7(a)) and the effective number of independent parameters (p_d) (3.7(b)) for models with the number of clusters k varying between 1 and $K = 10$. Figure 3.7(a) also provides the relative percentage of difference in DIC from k to $k+1$ clusters, where "-" means a decrease in DIC and "+" means an increase in DIC.

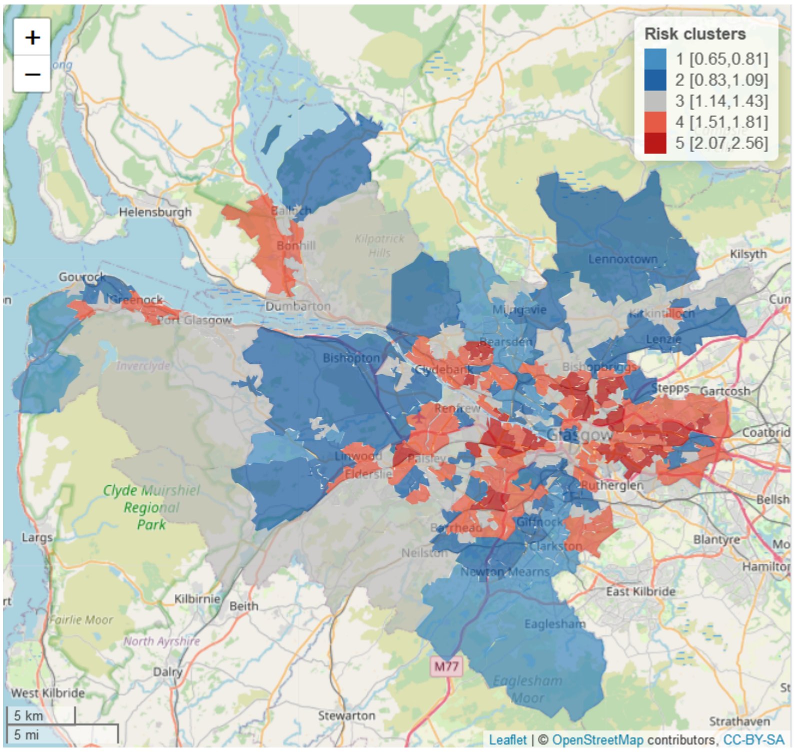


Figure 3.8: A map of the five estimated risk clusters from model P4 in Greater Glasgow in 2016. Values in square brackets show the minimum and maximum risk estimates in each risk cluster.

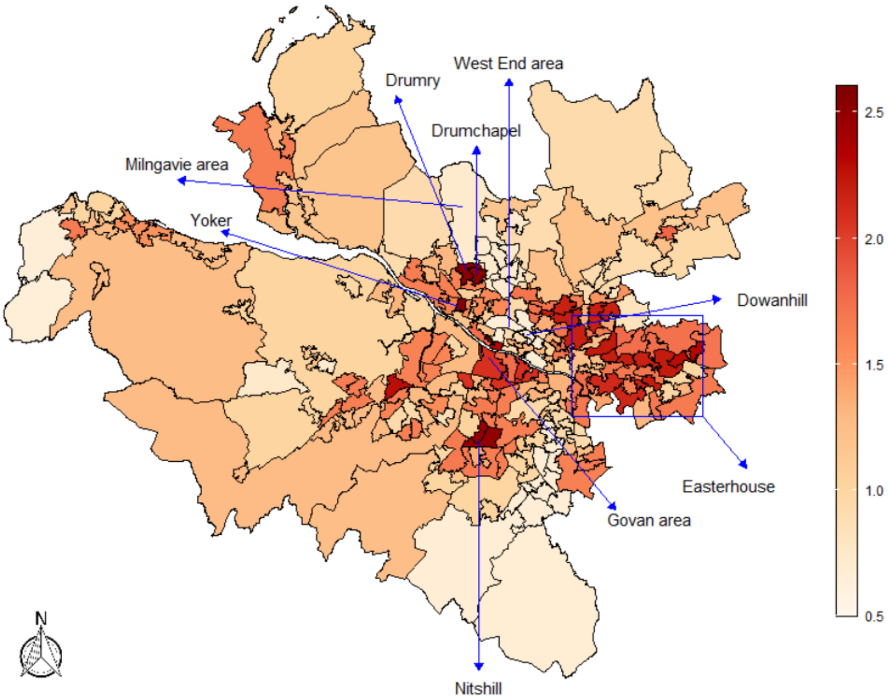


Figure 3.9: The estimated spatial risk pattern (posterior median) for respiratory disease in the Greater Glasgow and Clyde Health Board region in 2016 from the proposed model P4. The arrows in the map identify some typically high or low-risk areas.

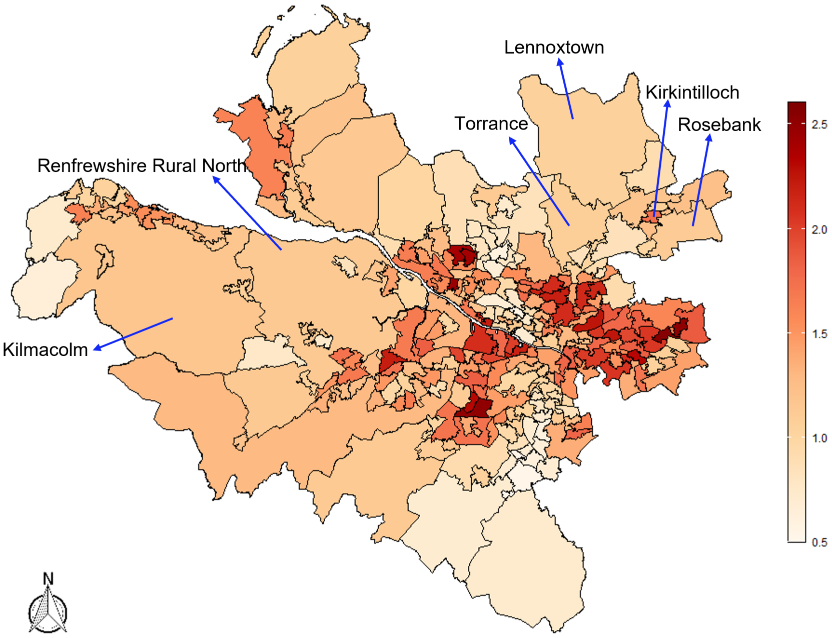


Figure 3.10: The estimated spatial risk pattern (posterior median) for respiratory disease in the Greater Glasgow and Clyde Health Board region in 2016 from the Leroux CAR model. The arrows in the map identify some neighbouring areas that have been identified as being in different clusters by model P4.

3.6 Discussion

The proposed methodology aims to achieve a better performance in capturing the spatial pattern of disease risk and identifying clusters of areas with high risks. Firstly, I use k-means clustering to partition all areal units into clusters based on the natural logarithm of the SIR, and the resulting candidate cluster structures are used to estimate a set of candidate neighbourhood matrices. Then separate spatial Bayesian hierarchical models are fitted to the data for each of the candidate matrices/cluster structures, and the choice of the best model is determined by a model selection rule. Four approaches are proposed in Section 3.3.3 to choose the best model; **Approach III** is inappropriate due to its great sensitivity to the threshold value. **Approach I** and **II** work by comparing the DIC values of the models with varying numbers of clusters, while **Approach IV** focuses on model parsimony and selects the model with the cluster structure having the smallest effective number of independent parameters.

The simulation study in Section 3.4 shows that the proposed models P1, P2 and P4 produce accurate risk estimates when there are clusters in the data ($Z > 0$), and particularly outperform the commonly used non-cluster Leroux model. This improved performance is likely because our models attempt to estimate an appropriate neighbourhood matrix from the data rather than naively using the border sharing $\boldsymbol{W}$, therefore avoiding the incorrect smoothing of the random effects between pairs of neighbouring areas that have very different disease risks. Models P1, P2 and P4 can also accurately identify clusters, with high ARI values being obtained in the presence of clusters. However when $Z = 0$, model P4 and the Leroux model perform comparably well and are better than models P1 and P2. The simulation study also suggests that model P4 performs well in terms of both risk estimation and cluster identification for diseases with moderate to large values of expected cases. However, when the number of expected cases is low (e.g. lower than 30 cases) and the cluster differences are small, the proposed models are less accurate in estimating disease risk and identifying the correct cluster structure. This is because in this scenario the clusters are hardest to identify based on their small size and small numbers of disease cases. The motivating application illustrates that overall the low-risk clusters are in the south of the city center (e.g. Clarkston, Newton Mearns) and also in the north such as Milngaive and Bearsden. In contrast, the high-risk clusters are mostly located in the east and west of the city, e.g. Easterhouse, Clydebank and Drumchapel. These results suggest that people living in the wealthier areas appear to be at lower risk of respiratory hospital admissions.

The methodology has a nature of estimating the neighbourhood matrix from the data so that the spatial autocorrelation is not always enforced between the random effects of neighbouring areas. The estimated neighbourhood matrix is a better representation of the spatial autocorrelation structure of the risk surface than the border sharing $\boldsymbol{W}$, which represents the correlation structure simply based on geographical adjacency. The methodology also has limitations. Firstly, it generates K fixed candidate neighbourhood matrices and does not quantify their uncertainty when estimating the risk surface. Secondly, the approach has to apply k-means clustering to the data K times in the first stage and then fit the spatial model K times in the model comparison procedure in stage two, which results in increased computational complexity especially when K is large. Furthermore, similarly to the approach proposed by Anderson et al. (2014), the generation of the the candidate cluster structures/neighbourhood matrices simply relies on a single clustering method (k-means clustering in our context), which means that for each fixed number of clusters only one spatial cluster structure is allowed as a candidate structure for the data. These issues will be explored and addressed in Chapter 4, where the spatial pattern in disease risk and the clusters of areas that exhibit elevated risks can be simultaneously estimated in a single model rather than by comparing multiple models as in this chapter. Another limitation to this work is that the clustering models proposed here are only compared to the non-clustering Leroux model in the simulation study, but they have not been compared to a method that accounts for risk clusters such as Anderson et al. (2014), Knorr-Held and Raßer (2000). This is because software to implement these complex estimation methods is not publicly available, and also because they use different modelling structures which may affect the results. However, such a comparison is very useful as it can assess the ability of our method to identify the correct cluster structure compared to other clustering models. Therefore, a fair comparison between the proposed methodology and another clustering approach would be worthwhile to consider in the future. Other avenues for future work include extending the spatial models to spatio-temporal models and constructing the candidate neighbourhood matrices using different definitions to that considered here. For example, the elements of the neighbourhood matrix relating to adjacent areas can be continuous and allowed to vary between 0 and 1, and in this way we can flexibly control the degree of smoothing between neighbouring random effects based on some geographical properties, such as the physical distance between pairs of neighbours.

Chapter 4

Estimating spatial disease risks and identifying clusters via clustering-based adjacency modelling

4.1 Introduction

In Chapter 3, a two-stage modelling approach is introduced to estimate disease risk and spatial clusters/discontinuities in the risk surface. In stage one k-means clustering is used to partition the entire study region into clusters of areas with different risk levels. With these clusters, the border sharing neighbourhood matrix $\boldsymbol{W}$ (i.e. $w_{ij} = 1$ if areas (i, j) share a common border in geography, otherwise, $w_{ij} = 0$) is locally altered so as to not have neighbours in different clusters. This creates a set of candidate $\boldsymbol{W}$ matrices based on differing numbers of clusters, which represent a range of possible cluster/discontinuity structures in the data. In stage two each of the candidate neighbourhood matrices is used to fit a separate Bayesian hierarchical model to the data. The most appropriate neighbourhood matrix corresponding to a given cluster structure is selected using model selection rules. This approach accounts for spatial clusters/discontinuities in the data by removing unnecessary borders between neighbouring areas through a clustering step, so that risks are not allowed to smooth towards their geographically neighbours that have significantly different risks in the modelling process. However, it has limitations such as the uncertainty in $\boldsymbol{W}$ not being measured and the candidate cluster structures all relying on k-means clustering method. To overcome these issues, an alternative two-stage approach is introduced in this chapter.

The methodology proposed here aims to estimate the spatial pattern of disease risk and identify spatial clusters/discontinuities in the risk surface by applying clustering to the data before the Bayesian modelling process, however, it differs from that used in the previous chapter in two main aspects. Firstly, instead of simply using k-means clustering, the approach here constructs a much bigger set of candidate cluster/discontinuity structures using a range of clustering methods, which gives much greater flexibility in cluster identification. Secondly, instead of choosing the best neighbourhood matrix via a comparison of multiple models, here the uncertainty in $\boldsymbol{W}$ is quantified by treating it as a random parameter in the modelling process, and the best choice of $\boldsymbol{W}$ is mainly informed by the data within a single model.

The remainder of this chapter is organised as follows. Section 4.2 outlines the proposed methodology. Section 4.3 examines its effectiveness against a commonly used model in the literature using simulated data. The sensitivity of the methodology to disease prevalence is assessed in Section 4.4. Section 4.5 applies the methodology to the respiratory disease data in the Greater Glasgow region in 2016. Finally, Section 4.6 further discusses the advantages of the approach and the future development.

4.2 Methodology

I propose a two-stage modelling approach for estimating the spatial pattern in disease risk and simultaneously detecting clusters of areas with elevated or reduced risks compared to their neighbours. The first stage generates a large collection of cluster structures from applying different clustering methods to the data of interest. For each clustering method, given a specified number of clusters k for the data the areal units are split into k clusters, where k is an integer ranging from 1 to K. From this, K cluster structures are generated and further used to produce K candidate neighbourhood matrices accordingly. In the second stage, unlike the approach in Chapter 3 where multiple models are fitted separately to each candidate cluster structure and the best cluster structure is chosen by a model comparison procedure, here a single Bayesian hierarchical model is fitted to the disease data, where the disease risks and the cluster structure implied by the best choice of $\boldsymbol{W}$ are estimated simultaneously.

4.2.1 Stage 1 — Generating candidate neighbourhood matrices via multiple clustering methods

A collection of candidate cluster structures for disease risk are estimated based on $c = 1,\ldots,M$ different clustering methods, and the $M = 8$ methods considered here are summarised in Table 4.1, which provides a key relating each clustering method to its corresponding value of c in our context. The methods include k-means (MacQueen et al., 1967) clustering, k-medoids (Park and Jun, 2009) clustering, hierarchical agglomerative (Hastie et al., 2009) clustering with centroid, complete, average and Ward's linkage, hierarchical divisive (Kaufman and Rousseeuw, 2009) clustering and expectation-maximisation (Fraley and Raftery, 2002) clustering. Note, hierarchical agglomerative clustering with single linkage is not considered due to its limitation of frequently suffering from the chaining effect (Yim and Ramdeen, 2015). In single linkage, the merge of two clusters simply depends on the smallest distance between one pair of data points, irrespective of others, therefore clusters can be too spread out and not compact enough. Besides, Anderson et al. (2014) have shown that single linkage exhibits poorer clustering performance than the other linkage methods. These clustering methods are applied to the data without regard to the spatial positions of the areal units, thus the clusters identified represent the number of different risk levels rather than the number of spatially distinct clusters, meaning that a single *"cluster"* will likely contain groups of areas that are not spatially connected.

Each clustering method c is used to compute $k = 1,\ldots,K$ distinct cluster structures, where structure k contains k clusters. Note, these K cluster structures are not necessarily nested, as for example a k-means solution with 4 clusters is not obtained by splitting one of the clusters from the k-means with 3 solutions into 2. The value K is chosen to be an upper limit for the number of clusters one would expect to find in the data, which must be specified by the user. As in Chapter 3, I set $K = 10$ as a conservative overly large choice, because this represents the number of distinct risk levels and not the number of spatially contiguous clusters. These candidate cluster structures are incorporated into the disease risk spatial model by specifying a set of candidate neighbourhood matrices, which means they relate to the random effects surface $\{\phi_i\}$ (as specified by the CAR prior, see Section 2.4.4). Therefore we initially estimate $\{\phi_i\}$ from the data and the general model (2.10) by

$$\tilde{\phi}_i = \ln\left(\frac{\mathbb{E}(Y_i)}{E_i}\right) - \boldsymbol{x}_i^\top \boldsymbol{\beta} \approx \ln\left(\frac{Y_i}{E_i}\right) - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}. \qquad (4.1)$$

This approximation replaces the unknown $\mathbb{E}(Y_i)$ with the observed data Y_i. Finally, the regression parameters are estimated for this initial stage assuming independence via maximum likelihood estimation, and are denoted above by $\hat{\boldsymbol{\beta}}$. These estimated spatial random effects $\{\tilde{\phi}_i\}$ are used to construct candidate cluster structures and corresponding neighbourhood matrices as described below.

Spatial clusters are obtained by applying each clustering method to $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_n)$. The cluster structure obtained from method c with k clusters is used to create a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k)}$ as follows:

$$w_{ij}^{(c,k)} = \begin{cases} 1, & \text{if areal units } (i,j) \text{ share a common border and are in the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Thus there is a one-to-one relationship between a candidate cluster structure and its corresponding neighbourhood matrix. Since $\left(\boldsymbol{W}^{(1,1)}, \boldsymbol{W}^{(2,1)}, \ldots, \boldsymbol{W}^{(M,1)}\right)$ all relate to the cluster structure with a single cluster containing all the areal units, they are the same and all equal to the border sharing $\boldsymbol{W}$ matrix, and this leads to $(K-1) \times M + 1$ candidate cluster structures in total. Altering the border sharing $\boldsymbol{W}$ in this way to allow for clusters means that if areas (i,j) share a border and are in the same cluster (i.e. have similar data values) then their random effects will be modelled as partially correlated and allowed to be smoothed towards each other (see equation (2.15)). In contrast, if they share a border but are in different clusters (i.e. have very different data values) then their random effects will be modelled as conditionally independent, thus not enforcing spatial smoothing between them.

Table 4.1: A key table for each value of c and its associated clustering method.

c	**clustering method**
$c=1$	k-means clustering (kmeans)
$c=2$	k-medoids clustering (kmedoids)
$c=3$	hierarchical agglomerative clustering with centroid linkage (agg_centroid)
$c=4$	hierarchical agglomerative clustering with complete linkage (agg_complete)
$c=5$	hierarchical agglomerative clustering with average linkage (agg_average)
$c=6$	hierarchical agglomerative clustering with ward linkage (agg_ward)
$c=7$	hierarchical divisive clustering (div)
$c=8$	expectation-maximisation clustering (EM)

4.2.2 Stage 2 — Bayesian spatial modelling

The second stage of the approach fits a model to the data that jointly estimates the spatial pattern in disease risk and an appropriate cluster/discontinuity structure. The latter is achieved by treating the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ as a parameter to be estimated from the set of candidates generated in stage 1. The overall proposed model is given by

$$\begin{aligned}
Y_i|E_i,R_i &\sim \text{Poisson}(E_iR_i), \quad i=1,\ldots,n,\\
\ln(R_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i, \qquad (4.3)\\
\beta_j &\sim \text{N}(0,1000), \text{ for } j=0,\ldots,p,\\
\boldsymbol{\phi} &\sim \text{N}\left(\boldsymbol{0}, \tau^2 \boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})^{-1}\right),\\
\widetilde{\boldsymbol{W}} &\sim \text{Discrete uniform}\left(\boldsymbol{W}^{(1,1)},\ldots,\boldsymbol{W}^{(1,K)},\boldsymbol{W}^{(2,2)},\ldots,\boldsymbol{W}^{(2,K)},\ldots,\boldsymbol{W}^{(M,2)},\ldots,\boldsymbol{W}^{(M,K)}\right),\\
\tau^2 &\sim \text{Inverse-Gamma}(1,0.01).
\end{aligned}$$

The spatial variation in disease risk is modelled by covariates $\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}$ and spatial random effects $\{\phi_i\}$. The spatial random effects $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_n)$ account for the residual variation in the data after the effects of the covariates have been removed, and are modelled by a multivariate Gaussian distribution $\boldsymbol{\phi} \sim \text{N}\left(\boldsymbol{0}, \tau^2 \boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})^{-1}\right)$. The spatially correlated precision matrix is $\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}}) = \rho(\text{diag}(\widetilde{\boldsymbol{W}}\boldsymbol{1}) - \widetilde{\boldsymbol{W}}) + (1-\rho)\boldsymbol{I}$, which corresponds to the Leroux CAR model (Leroux et al., 2000). The full conditional distribution of this CAR model for area i is given by

$$\phi_i|\boldsymbol{\phi}_{-i},\widetilde{\boldsymbol{W}} \sim \text{N}\left(\frac{\rho\sum_{j=1}^n w_{ij}\phi_j}{\rho\sum_{j=1}^n w_{ij}+1-\rho}, \frac{\tau^2}{\rho\sum_{j=1}^n w_{ij}+1-\rho}\right), \qquad (4.4)$$

where $\boldsymbol{\phi}_{-i} = (\phi_1,\ldots,\phi_{i-1},\phi_{i+1},\ldots,\phi_n)$. The conditional expectation of ϕ_i is a weighted average of the random effects in neighbouring areas, with binary weights based on the current choice of $\widetilde{\boldsymbol{W}}$ matrix. The parameter ρ controls the amount of spatial smoothing (correlation) globally across the study region, with values close to 1 corresponding to strong spatial autocorrelation while a value of zero corresponds to spatial independence. However, in our model the spatial autocorrelation structure is modelled locally for each pair of neighbouring areas by estimating an appropriate neighbourhood matrix for the data, which may make the estimation of a single global parameter redundant. In addition, as can be seen from equation (2.15), if ρ is estimated as 0 then two geographically adjacent random effects (ϕ_i,ϕ_j) will be conditionally independent even if areal units (i,j) are in the same cluster and so have $w_{ij}=1$. This is because if $\rho=0$ then $\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}}) = \boldsymbol{I}$ and hence $\widetilde{\boldsymbol{W}}$ no longer appears in the precision matrix. Therefore, I follow Lee and Mitchell (2012) and Lee and Mitchell (2013), and fix $\rho = 0.99$ in this model to enforce strong spatial autocorrelation globally, whose structure is then adjusted locally by estimating the neighbourhood matrix. $\rho = 1$ is not used because the approach could produce a candidate neighbourhood matrix where an areal unit has no neighbours due to it being a singleton cluster. This will cause $\sum_{j=1}^{n} w_{ij} = 0$ for the area i in question, which leads to an infinite mean and variance for ϕ_i from (4.4).

With ρ fixed at 0.99 in the analysis for this section, the partial correlation between (ϕ_i, ϕ_j) conditioning on the remaining effects $\boldsymbol{\phi}_{-ij}$ is given by

$$\begin{aligned}\text{Corr}(\phi_i, \phi_j | \boldsymbol{\phi}_{-ij}) &= \frac{\rho w_{ij}}{\sqrt{\left(\rho \sum_{v=1}^{n} w_{iv} + 1 - \rho\right)\left(\rho \sum_{v=1}^{n} w_{jv} + 1 - \rho\right)}} \\ &= \frac{0.99 w_{ij}}{\sqrt{\left(0.99 \sum_{v=1}^{n} w_{iv} + 1 - 0.99\right)\left(0.99 \sum_{v=1}^{n} w_{jv} + 1 - 0.99\right)}}.\end{aligned}$$

Therefore, (ϕ_i, ϕ_j) are only partially correlated if $w_{ij} = 1$ (i.e. areas (i, j) are adjacent and in the same cluster), otherwise, they are modelled as conditionally independent as $\text{Corr}(\phi_i, \phi_j | \boldsymbol{\phi}_{-ij}) = 0$. Hence $\widetilde{\boldsymbol{W}}$ determines the spatial correlation structure imposed by the model.

Each covariate regression parameter β_j is assigned an independent weakly informative zero-mean Gaussian prior distribution with a large variance, to ensure its value is mainly informed by the data. The conjugate Inverse-Gamma prior, Inverse-Gamma$(1, 0.01)$, is specified for τ^2 to allow the parameter to be updated via Gibbs sampling. Since there is no prior knowledge about the number of clusters present in the data, $\widetilde{\boldsymbol{W}}$ is assigned a discrete uniform prior distribution whose possible values are the set of candidates corresponding to the cluster structures estimated in stage 1. The clustering stage elicits multiple candidate neighbourhood matrices that are equal to the border sharing $\boldsymbol{W}$, which occur when the number of clusters $k = 1$ for each clustering method. Therefore only one of these is included in the discrete uniform prior to avoid the border sharing $\boldsymbol{W}$ being given a larger prior weight compared to the other candidate values. We include this border sharing $\boldsymbol{W}$ in the model because it corresponds to a globally spatially smooth risk surface with no clusters. To achieve identifiability, the spatial random effects are zero-mean centred.

4.2.3 Inference

The posterior summaries of each parameter are obtained by drawing samples from the posterior distribution using Markov chain Monte Carlo (MCMC) simulation, which combines Metropolis-Hastings (Metropolis et al., 1953, Hastings, 1970) and Gibbs sampling steps (Geman and Geman, 1984). The MCMC algorithm is written (as part of this thesis) and implemented in R (R Core Team, 2013). In order to speed up computation, the updates of the random effects are written in the more efficient language C++ via the R package Rcpp (Eddelbuettel et al., 2011, Eddelbuettel, 2013). In addition, as $\widetilde{\boldsymbol{W}}$ is a large but sparse matrix I exploit its triplet form to improve computational efficiency. Inference is based on 300,000 MCMC samples with a burn-in period of 200,000. The chain is thinned by 10, due to limited computer memory and to make the samples closer to be independent, and so the posterior estimates are based on a total of 10,000 samples. Convergence is assessed by checking parameter trace plots. Details of each step of the MCMC sampler for model (4.3) are as follows.

<u>Update $\boldsymbol{\beta}$</u>

The full conditional distribution for each β_j is

$$\begin{aligned} f(\beta_j|\boldsymbol{Y}) &\propto \prod_{i=1}^{n} \text{Poisson}(Y_i|\beta_j) \times \text{N}(\beta_j|0,1000) \\ &\propto \prod_{i=1}^{n} \left(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i)\right)^{Y_i} \exp\left(-E_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i)\right) \times \exp\left(\frac{-\beta_j^2}{2000}\right). \end{aligned}$$

β_j is updated via the Metropolis-Hastings algorithm, with a proposal β_j^* randomly sampled from the proposal distribution $\beta_j^* \sim \text{N}(\beta_j^c, v_{\beta_j})$, where β_j^c is the current value of β_j. The proposal variance v_{β_j} can be altered within the algorithm to keep an acceptance rate between 40% and 60% for the parameter of low dimension (Gelman et al., 1996). The acceptance probability of moving from β_j^c to β_j^* is given by $\min\left\{1, \frac{f(\beta_j^*|\boldsymbol{Y})}{f(\beta_j^c|\boldsymbol{Y})}\right\}$.

<u>Update $\boldsymbol{\phi}$</u>

The full conditional distribution for ϕ_i is

$$\begin{aligned} f(\phi_i|Y_i) &\propto \text{Poisson}(Y_i|\phi_i) \times \text{N}(\phi_i|\boldsymbol{\phi}_{-i}) \\ &\propto \left(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i)\right)^{Y_i} \exp\left(-E_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i)\right) \times \\ &\quad \text{N}\left(\frac{0.99 \sum_{j=1}^{n} w_{ij}\phi_j}{0.99 \sum_{j=1}^{n} w_{ij} + 1 - 0.99}, \frac{\tau^2}{0.99 \sum_{j=1}^{n} w_{ij} + 1 - 0.99}\right). \end{aligned}$$

Each ϕ_i is updated individually via the Metropolis-Hastings algorithm, with a proposal ϕ_i^* randomly sampled from the proposal distribution $\phi_i^* \sim \text{N}(\phi_i^c, v_{\phi_i})$, where ϕ_i^c is the current value of ϕ_i. The proposal variance v_{ϕ_i} can be altered within the algorithm to maintain an acceptance rate between 40% and 60%. The acceptance probability of ϕ_i^* is given by $\min\left\{1, \frac{f(\phi_i^*|Y_i)}{f(\phi_i^c|Y_i)}\right\}$.

Update τ^2

The full conditional distribution for τ^2 is

$$\begin{aligned}
f(\tau^2|\boldsymbol{\phi}) &\propto \text{N}\left(\boldsymbol{\phi}|\mathbf{0}, \tau^2\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})^{-1}\right) \times \text{Inverse-Gamma}(1, 0.01)\\
&\propto \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})\boldsymbol{\phi})\right) \times \left(\frac{1}{\tau^2}\right)^{(1+\frac{n}{2}+1)} \exp\left(-\frac{0.01}{\tau^2}\right)\\
&\propto \left(\frac{1}{\tau^2}\right)^{(1+\frac{n}{2}+1)} \exp\left(-\frac{1}{\tau^2}\left(0.01+\frac{1}{2}\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})\boldsymbol{\phi}\right)\right)\\
&\sim \text{Inverse-Gamma}(a,b),
\end{aligned}$$

where $a = 1+\frac{n}{2}$ and $b = 0.01+\frac{1}{2}\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})\boldsymbol{\phi}$. Here $\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}}) = 0.99(\text{diag}(\widetilde{\boldsymbol{W}}\mathbf{1})-\widetilde{\boldsymbol{W}}) + (1-0.99)\boldsymbol{I}$ as $\rho = 0.99$ in the model. τ^2 is evaluated at each iteration of Gibbs sampling by directly drawing samples from the above Inverse-Gamma distribution.

Update $\widetilde{\boldsymbol{W}}$

The full conditional distribution for $\widetilde{\boldsymbol{W}}$ is

$$\begin{aligned}
f(\widetilde{\boldsymbol{W}}|\boldsymbol{\phi}) &\propto \text{N}\left(\boldsymbol{\phi}|\mathbf{0}, \tau^2\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})^{-1}\right) \times f\left(\widetilde{\boldsymbol{W}} = \boldsymbol{W}^{(c,k)}\right)\\
&\propto ||\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})||^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\widetilde{\boldsymbol{W}})\boldsymbol{\phi})\tau^{-2}\right),
\end{aligned}$$

where $||\cdot||$ denotes the determinant of a matrix. The set of potential $\widetilde{\boldsymbol{W}}$ matrices, $(\boldsymbol{W}^{(1,1)},\ldots,\boldsymbol{W}^{(1,K)},\boldsymbol{W}^{(2,2)},\ldots,\boldsymbol{W}^{(2,K)},\ldots,\boldsymbol{W}^{(M,2)},\ldots,\boldsymbol{W}^{(M,K)})$, are generated in stage one and the Bayesian model in stage two selects which of these candidate matrices, corresponding to a given cluster structure, is the most appropriate for the data. A Metropolis-Hastings algorithm is used to update the choice of $\widetilde{\boldsymbol{W}}$, and two approaches are considered for this. **Approach 1** updates $\widetilde{\boldsymbol{W}}$ by using a Metropolis-Hastings step consisting of two MCMC moves, which are outlined below.

1. If the current value of $\widetilde{\boldsymbol{W}}$ in the Markov chain is $\boldsymbol{W}^{(c,k)}$, then a new value $\boldsymbol{W}^{(c,l)}$ is proposed uniformly from the set of candidate matrices

$\left(\boldsymbol{W}^{(c,k-s)},\ldots,\boldsymbol{W}^{(c,k-1)},\boldsymbol{W}^{(c,k+1)},\ldots,\boldsymbol{W}^{(c,k+s)}\right)$, which correspond to the candidate cluster structures generated from the same clustering method as the current choice but with a different number of clusters. Here s is a parameter controlling the acceptance rates and mixing of the update, and exploratory model runs suggested that $s = 2$ leads to good estimation performance. If k is close to 1 or $K(= 10)$, some of the theoretical proposal matrices are likely not to exist in practice, as a result, the number of proposal matrices is reduced and the associated probabilities need to be adjusted accordingly. Since the proposal distribution is likely to be asymmetric, the acceptance probability of $\boldsymbol{W}^{(c,l)}$ is given by $\min\left\{1,\frac{f(\boldsymbol{W}^{(c,l)}|\boldsymbol{Y})/g\left(\boldsymbol{W}^{(c,l)}|\ \boldsymbol{W}^{(c,k)}\right)}{f(\boldsymbol{W}^{(c,k)}|\boldsymbol{Y})/g\left(\boldsymbol{W}^{(c,k)}|\boldsymbol{W}^{(c,l)}\right)}\right\}$, where $g(\boldsymbol{W}^{(c,l)}|\ \boldsymbol{W}^{(c,k)})$ is the probability of proposing $\boldsymbol{W}^{(c,l)}$ given that the current choice is $\boldsymbol{W}^{(c,k)}$.

2. If the current value of $\widetilde{\boldsymbol{W}}$ after the first move is $\boldsymbol{W}^{(c,k')}$, then a new proposal $\boldsymbol{W}^{(h,k')}$ is drawn uniformly from the set $\left(\boldsymbol{W}^{(1,k')},\ldots,\boldsymbol{W}^{(c-1,k')},\boldsymbol{W}^{(c+1,k')},\ldots,\boldsymbol{W}^{(M,k')}\right)$, which correspond to the candidate cluster structures that have the same number of clusters as the current choice but are generated from a different clustering method. Since the proposal distribution is symmetric, the acceptance probability of moving from $\boldsymbol{W}^{(c,k')}$ to $\boldsymbol{W}^{(h,k')}$ is calculated by $\min\left\{1,\frac{f(\boldsymbol{W}^{(h,k')}|\boldsymbol{Y})}{f(\boldsymbol{W}^{(c,k')}|\boldsymbol{Y})}\right\}$.

Table 4.2 provides an intuitive example illustrating the two MCMC moves under **Approach 1**. Suppose $\boldsymbol{W}^{(3,5)}$ is the current choice of $\widetilde{\boldsymbol{W}}$, then the candidate matrices highlighted in blue comprise the sample space from which a proposal matrix is drawn in the first move. If the current value of $\widetilde{\boldsymbol{W}}$ is $\boldsymbol{W}^{(3,3)}$ after the first move, then the candidate matrices highlighted in red comprise the sample space for proposing a new matrix in the second move.

Table 4.2: An example that illustrates the two MCMC moves under **Approach 1**.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=K$
$c=1$	$\boldsymbol{W}^{(1,1)}$	$\boldsymbol{W}^{(1,2)}$	$\boldsymbol{W}^{(1,3)}$	$\boldsymbol{W}^{(1,4)}$	$\boldsymbol{W}^{(1,5)}$	$\boldsymbol{W}^{(1,6)}$	$\boldsymbol{W}^{(1,7)}$	$\boldsymbol{W}^{(1,8)}$	$\boldsymbol{W}^{(1,9)}$	$\boldsymbol{W}^{(1,10)}$
$c=2$	$\boldsymbol{W}^{(2,1)}$	$\boldsymbol{W}^{(2,2)}$	$\boldsymbol{W}^{(2,3)}$	$\boldsymbol{W}^{(2,4)}$	$\boldsymbol{W}^{(2,5)}$	$\boldsymbol{W}^{(2,6)}$	$\boldsymbol{W}^{(2,7)}$	$\boldsymbol{W}^{(2,8)}$	$\boldsymbol{W}^{(1,9)}$	$\boldsymbol{W}^{(1,10)}$
$c=3$	$\boldsymbol{W}^{(3,1)}$	$\boldsymbol{W}^{(3,2)}$	$\boldsymbol{W}^{(3,3)}$	$\boldsymbol{W}^{(3,4)}$	$\boldsymbol{W}^{(3,5)}$	$\boldsymbol{W}^{(3,6)}$	$\boldsymbol{W}^{(3,7)}$	$\boldsymbol{W}^{(3,8)}$	$\boldsymbol{W}^{(3,9)}$	$\boldsymbol{W}^{(3,10)}$
$c=4$	$\boldsymbol{W}^{(4,1)}$	$\boldsymbol{W}^{(4,2)}$	$\boldsymbol{W}^{(4,3)}$	$\boldsymbol{W}^{(4,4)}$	$\boldsymbol{W}^{(4,5)}$	$\boldsymbol{W}^{(4,6)}$	$\boldsymbol{W}^{(4,7)}$	$\boldsymbol{W}^{(4,8)}$	$\boldsymbol{W}^{(4,9)}$	$\boldsymbol{W}^{(4,10)}$
$c=5$	$\boldsymbol{W}^{(5,1)}$	$\boldsymbol{W}^{(5,2)}$	$\boldsymbol{W}^{(5,3)}$	$\boldsymbol{W}^{(5,4)}$	$\boldsymbol{W}^{(5,5)}$	$\boldsymbol{W}^{(5,6)}$	$\boldsymbol{W}^{(5,7)}$	$\boldsymbol{W}^{(5,8)}$	$\boldsymbol{W}^{(5,9)}$	$\boldsymbol{W}^{(5,10)}$
$c=6$	$\boldsymbol{W}^{(6,1)}$	$\boldsymbol{W}^{(6,2)}$	$\boldsymbol{W}^{(6,3)}$	$\boldsymbol{W}^{(6,4)}$	$\boldsymbol{W}^{(6,5)}$	$\boldsymbol{W}^{(6,6)}$	$\boldsymbol{W}^{(6,7)}$	$\boldsymbol{W}^{(6,8)}$	$\boldsymbol{W}^{(6,9)}$	$\boldsymbol{W}^{(6,10)}$
$c=7$	$\boldsymbol{W}^{(7,1)}$	$\boldsymbol{W}^{(7,2)}$	$\boldsymbol{W}^{(7,3)}$	$\boldsymbol{W}^{(7,4)}$	$\boldsymbol{W}^{(7,5)}$	$\boldsymbol{W}^{(7,6)}$	$\boldsymbol{W}^{(7,7)}$	$\boldsymbol{W}^{(7,8)}$	$\boldsymbol{W}^{(7,9)}$	$\boldsymbol{W}^{(7,10)}$
$c=8$	$\boldsymbol{W}^{(8,1)}$	$\boldsymbol{W}^{(8,2)}$	$\boldsymbol{W}^{(8,3)}$	$\boldsymbol{W}^{(8,4)}$	$\boldsymbol{W}^{(8,5)}$	$\boldsymbol{W}^{(8,6)}$	$\boldsymbol{W}^{(8,7)}$	$\boldsymbol{W}^{(8,8)}$	$\boldsymbol{W}^{(8,9)}$	$\boldsymbol{W}^{(8,10)}$

Approach 2 updates $\widetilde{\boldsymbol{W}}$ by a Metropolis-Hastings step which only proposes a new neighbourhood matrix once from a specified sample space for $\widetilde{\boldsymbol{W}}$ at each MCMC iteration. The set of candidate matrices are ordered and denoted by $(\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_{M\times K})$,

where $\boldsymbol{\Lambda}_{(c-1)\times K+k}$ represents matrix $\boldsymbol{W}^{(c,k)}$. We propose a matrix from a maximum of s-proximity of the current matrix. If the current choice is $\boldsymbol{W}^{(c,k)}$, a proposal is drawn from $\left(\boldsymbol{\Lambda}_{(c-1)\times K+k-s},\ldots,\boldsymbol{\Lambda}_{(c-1)\times K+k-1},\boldsymbol{\Lambda}_{(c-1)\times K+k+1},\ldots,\boldsymbol{\Lambda}_{(c-1)\times K+k+s}\right)$ with equal probability of selecting each matrix, and likewise $s = 2$ is used here. When $\boldsymbol{W}^{(c,k)}$ is close to an endpoint (i.e. either $\boldsymbol{\Lambda}_1$ or $\boldsymbol{\Lambda}_{M\times K}$), some of these theoretical proposal matrices may not exist in practice, hence the number of proposal matrices is reduced and the associated probabilities need to be adjusted accordingly. Since the proposal distribution is likely to be asymmetric, the acceptance probability of a move from $\boldsymbol{W}^{(c,k)}$ to $\boldsymbol{W}^*$ is calculated by $\min\left\{1, \frac{f(\boldsymbol{W}^*|\boldsymbol{Y})/g\left(\boldsymbol{W}^*|\ \boldsymbol{W}^{(c,k)}\right)}{f(\boldsymbol{W}^{(c,k)}|\boldsymbol{Y})/g\left(\boldsymbol{W}^{(c,k)}|\boldsymbol{W}^*\right)}\right\}$, where $g\left(\boldsymbol{W}^*|\ \boldsymbol{W}^{(c,k)}\right)$ is the probability of proposing $\boldsymbol{W}^*$ given that the current matrix is $\boldsymbol{W}^{(c,k)}$.

Since $\widetilde{\boldsymbol{W}}$ follows a discrete distribution and the number of clusters requires an integer value, the posterior mode value from the posterior distribution of $\widetilde{\boldsymbol{W}}$, representing the most likely occurring cluster structure (or neighbourhood matrix) across all the MCMC samples, can be used to estimate the optimal cluster/discontinuity structure. In contrast, the remaining parameters are summarised by their posterior medians. Additionally, pilot runs also suggested that the MCMC moves of $\widetilde{\boldsymbol{W}}$ are more likely to be accepted if their corresponding candidate cluster structures have higher similarity in terms of the adjusted Rand Index. If the candidate cluster structures and neighbourhood matrices are very different from each other, the moves of $\widetilde{\boldsymbol{W}}$ are less easy and longer MCMC runs may be needed for model convergence.

4.3 Simulation Study

4.3.1 Aim

A simulation study is carried out to assess the performance of the proposed methodology. The study compares the models where $\widetilde{\boldsymbol{W}}$ is updated using **Approach 1** or **Approach 2** in the MCMC sampler as outlined in Section 4.2.3. Their performances are then compared against the Leroux model (Leroux et al., 2000) described in Section 2.4.4.

4.3.2 Data generation

Clustered disease data are generated in the same way as described in Chapter 3. The template for the study region is the set of 257 Intermediate Zones comprising the Greater Glasgow and Clyde Health Board, which is shown in Figure 3.3. The simulated data consists of three clusters of areas with a high, medium and low level of disease risk, and are generated via the Poisson model (3.5) outlined in Chapter 3. The expected disease counts $\boldsymbol{E}$ quantify the disease prevalence and are set equal to 100 for the analyses described in this section. The extent of the differences between clusters, that is the magnitude of the differences between the three risk levels, is controlled by multiplying the piecewise constant mean values $\boldsymbol{\mu} = \{-1, 0, 1\}$ for $\boldsymbol{\phi}$ by a constant scalar Z before generating the data. Larger values of Z represent larger differences in disease risk between the clusters, which should thus be easier to identify. Values of $Z = 1, 0.5, 0$ are used in this study; $Z = 1$ indicates large differences between the clusters, $Z = 0.5$ corresponds to a case where there are smaller differences in the spatial surface, and $Z = 0$ corresponds to a spatially smooth risk surface with no clusters, thus in this case one would expect to identify a single cluster covering the entire study region.

4.3.3 Results

One hundred data sets are simulated for each of the three scenarios ($Z = 1, 0.5, 0$). The proposed models that respectively use **Approach 1** and **Approach 2** to update the choice of $\widetilde{\boldsymbol{W}}$ within the MCMC sampler are compared against the Leroux model with the border sharing $\boldsymbol{W}$ applied, which is commonly used in disease mapping. The results of the study over all simulated data sets are summarised in Figures 4.1, 4.2, 4.3 and 4.4, and outlined in Table 4.3. The accuracy of the risk estimation for each modelling approach is quantified by the bias, root mean square error (RMSE) and 95% coverage probabilities for the overall risk estimates. The correctness of the identified cluster structure is measured by both the number of clusters estimated and the adjusted Rand Index between the true and estimated cluster structures. The adjusted Rand Index (ARI) proposed by Hubert and Arabie (1985) is a measure of the similarity between two cluster structures. A value of 1 indicates complete agreement between the two cluster structures, a value of 0 indicates that the data points are randomly allocated to the two cluster structures, and a value less than 0 indicates that the level of agreement between the two cluster structures is smaller than that arising from randomly allocated data points. For more information on the ARI, see Section 2.7.3.

Figures 4.1 and 4.2 respectively summarise the estimated $\widetilde{\boldsymbol{W}}$ over the 100 simulated data sets for models **Approach 1** and **Approach 2** for each scenario of Z. In the figures, each grid square represents a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k)}$ corresponding to a distinct cluster structure, where the horizontal axis denotes the number of clusters k and the vertical axis denotes the clustering method c. Note that the grid square on the bottom left corner corresponds to the border sharing $\boldsymbol{W} = \boldsymbol{W}^{(1,1)}, \ldots, \boldsymbol{W}^{(M,1)}$ (i.e. $k = 1$) which represents no clusters in disease risk. The figures also provide the percentage of times that each candidate neighbourhood matrix is identified as the most appropriate choice of $\widetilde{\boldsymbol{W}}$ (i.e. posterior mode) over all simulated data sets, with darker shading corresponding to a higher chance (probability) of being the best $\widetilde{\boldsymbol{W}}$.

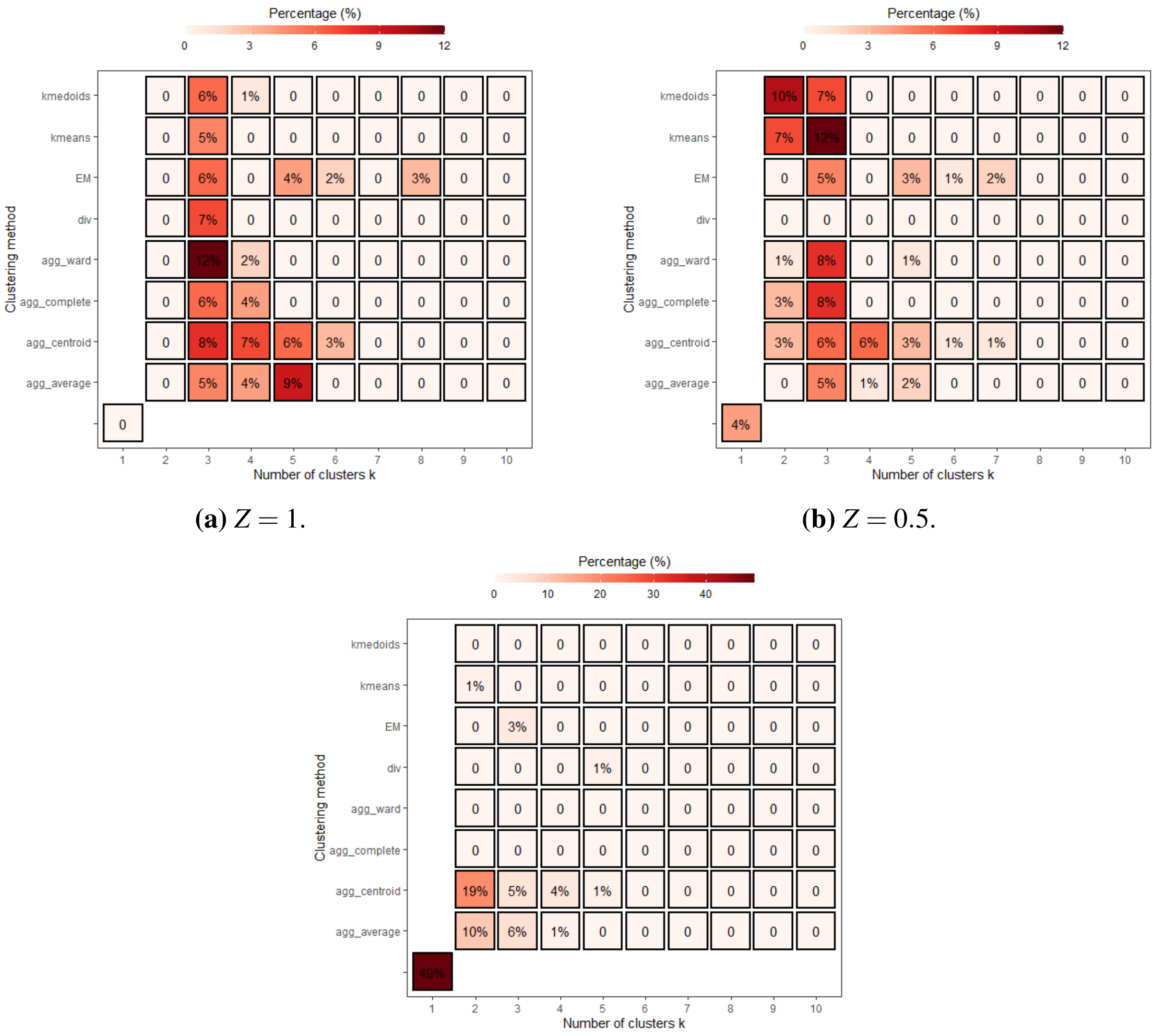
The results show that when the differences between the clusters are large in size ($Z = 1$), both modelling **Approach 1** and **Approach 2** estimate the correct number of clusters in the majority of the simulations. **Approach 1** does not show an underestimation of the number of clusters present, while **Approach 2** underestimates the number of clusters as 1 or 2 clusters in about 5% of the simulations. In the scenario where $Z = 0.5$, underestimation is more likely for both approaches because the differences between the true clusters are not very pronounced so that the clusters are more likely to be incorrectly joined together. In the scenario where $Z = 0$, models **Approach 1** and **Approach 2** correctly estimate the cluster structure, which is a single cluster covering the entire study region, in about 49% and 60% of the simulations respectively.

The top panel in Figure 4.3 displays boxplots of the number of clusters identified by each model and scenario, where the true values of 3 for $Z = 1, 0.5$ and 1 for $Z = 0$ are represented by dashed lines. The bottom panel shows boxplots of the adjusted Rand Index (ARI) values between the true and estimated cluster structures over all simulated data sets. When $Z = 1$, both modelling **Approach 1** and **Approach 2** identify the correct number of clusters for the majority of data sets with a median of 3 clusters. The two models perform well in cluster identification with ARI values close to 1, while **Approach 1** has a lower standard deviation (0.040) compared with **Approach 2** (0.179). Although there are a number of simulations where the number of clusters is overestimated, they generally obtain high ARI values, suggesting that the clustering performance is not substantially affected and the estimated cluster structures overall match with the true cluster structure well even under slight overestimation.

When $Z = 0.5$, the differences between the clusters in disease risk are less pronounced and hence it is more difficult to identify the true cluster structure. As a result, the ARI values are on average lower than those obtained under $Z = 1$ and underestimation is slightly more likely for $Z = 0.5$. The unclear boundaries between clusters also explain the slightly increased number of data sets estimating inaccurate number of clusters. The median number of clusters identified by both modelling approaches is 3, but **Approach 1** has a lower standard deviation of 1.222 compared to 1.545 for **Approach 2**. Additionally, **Approach 1** produces higher ARI values, with a median of 0.906 compared to 0.806 for **Approach 2**. However, when there are no clusters in disease risk ($Z = 0$), **Approach 1** overestimates the number of clusters with a median value of 2 clusters in 51% of the simulations, which unsurprisingly leads to the very poor performance in terms of cluster identification with a median ARI of 0. We note that in the case where $Z = 0$, both approaches tend to overestimate the number of clusters. This is probably due to the independent random variation induced into the count data by generating $\{Y_i\}$ from the Poisson model. Although there are no artificially constructed clusters present, it does not necessarily mean that the simulated disease data have zero difference and clusters do not exist. The random variation can cause extra unintended clusters not classified as "true". Therefore, the two modelling approaches may partition a theoretically flat risk surface into clusters based on the seeming differences caused by sampling variation but not the actual risk differences from the spatial random effects.

Figure 4.4 displays a comparison of the relative performances of the two modelling approaches proposed here and the Leroux model which is fitted based on the border sharing $\boldsymbol{W}$. In the scenarios where $Z = 1, 0.5$, both models **Approach 1** and **Approach 2** outperform the Leroux model in terms of lower RMSE values and negligible bias of risk estimates close to zero, with reductions in the median RMSE respectively being 29.6% ($Z = 1$) and 19.0% ($Z = 0.5$) for **Approach 1**, and being 27.8% ($Z = 1$) and 15.0% ($Z = 0.5$) for **Approach 2**. This is probably because the Leroux model allows for incorrect smoothing of the random effects between clusters, which is not enforced by the proposed models. All three models exhibit good coverage ability since the coverages are close to the nominal 0.95 levels. **Approach 1** performs marginally better than **Approach 2** for providing a bit more precise risk estimates. When $Z = 0$, the three models generally perform well with the bias close to 0 and the coverage probabilities close to 0.95, while the Leroux model performs the best of the three in terms of RMSE, with a median of 0.008 compared with 0.023 for **Approach 1** and 0.019 for **Approach 2** respectively. The less accurate risk estimation of the proposed approaches might result from the poor cluster identification for $Z = 0$. Figure 4.3 shows that both approaches tend to overestimate the number of clusters when $Z = 0$, and this overestimation comes from a split in a true cluster. As touched on previously, the random variation induced into the simulated count data by the Poisson data model could make the risk surface less smooth than it is designed, thus the proposed approaches with a clustering mechanism are attempting to find non-smooth patterns based on the random variation and so are very likely to incorrectly split the theoretically zero-difference risks into different clusters. These inaccurate clusters lead to false smoothing between neighbours and thus reduce the modelling accuracy. It appears that the proposed approaches have slightly higher standard deviations in the performance metrics than the Leroux model. This is likely to be because they treat the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ as a parameter to be estimated from the data, however, in the Leroux model the neighbourhood matrix $\boldsymbol{W}$ is fixed based on the sharing a common border specification, thus the uncertainty in $\boldsymbol{W}$ is not measured in the modelling process, which results in less variability.

In summary, when there are clusters present in the data ($Z = 1, 0.5$), both modelling **Approach 1** and **Approach 2** generally outperform the Leroux model in terms of risk estimation. **Approach 1** provides better risk estimates and also identifies more accurate cluster structures than **Approach 2** especially for the scenario of $Z = 0.5$. When there is a spatially smooth risk surface ($Z = 0$), the two approaches tend to overestimate the number of clusters and provide less accurate risk estimates than the Leroux model.



Clustering method	1	2	3	4	5	6	7	8	9	10
kmedoids		0	6%	1%	0	0	0	0	0	0
kmeans		0	5%	0	0	0	0	0	0	0
EM		0	6%	0	4%	2%	0	3%	0	0
div		0	7%	0	0	0	0	0	0	0
agg_ward		0	12%	2%	0	0	0	0	0	0
agg_complete		0	6%	4%	0	0	0	0	0	0
agg_centroid		0	8%	7%	6%	3%	0	0	0	0
agg_average		0	5%	4%	9%	0	0	0	0	0
	0									

(a) $Z = 1$.

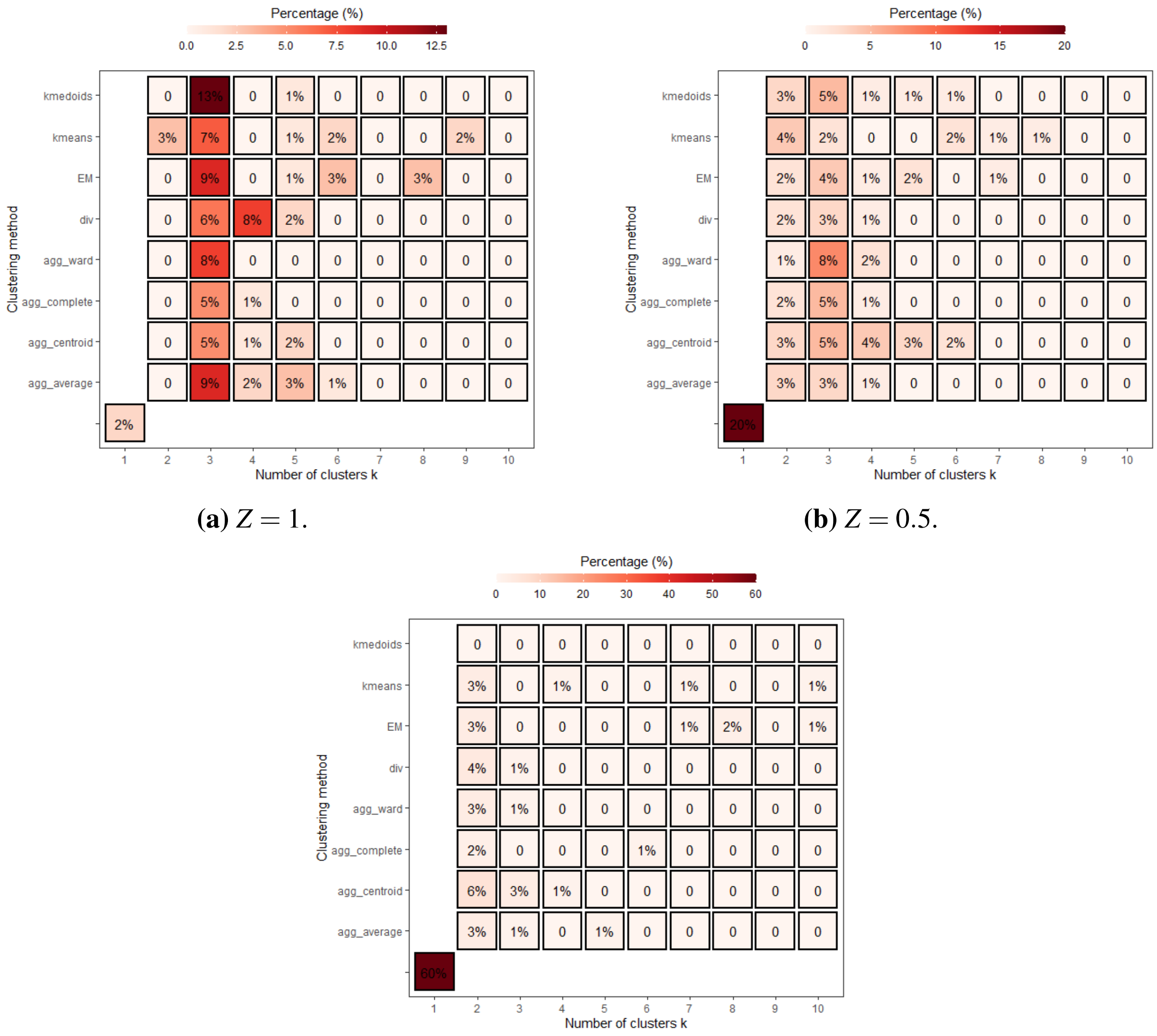
Clustering method	1	2	3	4	5	6	7	8	9	10
kmedoids		10%	7%	0	0	0	0	0	0	0
kmeans		7%	12%	0	0	0	0	0	0	0
EM		0	5%	0	3%	1%	2%	0	0	0
div		0	0	0	0	0	0	0	0	0
agg_ward		1%	8%	0	1%	0	0	0	0	0
agg_complete		3%	8%	0	0	0	0	0	0	0
agg_centroid		3%	6%	6%	3%	1%	1%	0	0	0
agg_average		0	5%	1%	2%	0	0	0	0	0
	4%									

(b) $Z = 0.5$.

Clustering method	1	2	3	4	5	6	7	8	9	10
kmedoids		0	0	0	0	0	0	0	0	0
kmeans		1%	0	0	0	0	0	0	0	0
EM		0	3%	0	0	0	0	0	0	0
div		0	0	0	1%	0	0	0	0	0
agg_ward		0	0	0	0	0	0	0	0	0
agg_complete		0	0	0	0	0	0	0	0	0
agg_centroid		19%	5%	4%	1%	0	0	0	0	0
agg_average		10%	6%	1%	0	0	0	0	0	0
	49%									

(c) $Z = 0$.

Figure 4.1: Summary of the estimated $\widetilde{\boldsymbol{W}}$ over 100 simulated data sets from the proposed modelling **Approach 1**. Each grid square shows the percentage of times that each candidate neighbourhood matrix is identified as the most appropriate choice of $\widetilde{\boldsymbol{W}}$ (i.e. the posterior mode) over 100 simulations for each scenario of Z.



	1	2	3	4	5	6	7	8	9	10
kmedoids		0	13%	0	1%	0	0	0	0	0
kmeans		3%	7%	0	1%	2%	0	0	2%	0
EM		0	9%	0	1%	3%	0	3%	0	0
div		0	6%	8%	2%	0	0	0	0	0
agg_ward		0	8%	0	0	0	0	0	0	0
agg_complete		0	5%	1%	0	0	0	0	0	0
agg_centroid		0	5%	1%	2%	0	0	0	0	0
agg_average		0	9%	2%	3%	1%	0	0	0	0
	2%									

(a) $Z = 1$.

	1	2	3	4	5	6	7	8	9	10
kmedoids		3%	5%	1%	1%	1%	0	0	0	0
kmeans		4%	2%	0	0	2%	1%	1%	0	0
EM		2%	4%	1%	2%	0	1%	0	0	0
div		2%	3%	1%	0	0	0	0	0	0
agg_ward		1%	8%	2%	0	0	0	0	0	0
agg_complete		2%	5%	1%	0	0	0	0	0	0
agg_centroid		3%	5%	4%	3%	2%	0	0	0	0
agg_average		3%	3%	1%	0	0	0	0	0	0
	20%									

(b) $Z = 0.5$.

	1	2	3	4	5	6	7	8	9	10
kmedoids		0	0	0	0	0	0	0	0	0
kmeans		3%	0	1%	0	0	1%	0	0	1%
EM		3%	0	0	0	0	1%	2%	0	1%
div		4%	1%	0	0	0	0	0	0	0
agg_ward		3%	1%	0	0	0	0	0	0	0
agg_complete		2%	0	0	0	1%	0	0	0	0
agg_centroid		6%	3%	1%	0	0	0	0	0	0
agg_average		3%	1%	0	1%	0	0	0	0	0
	60%									

(c) $Z = 0$.

Figure 4.2: Summary of the estimated $\widetilde{\boldsymbol{W}}$ over 100 simulated data sets from the proposed modelling **Approach 2**. Each grid square shows the percentage of times that each candidate neighbourhood matrix is identified as the most appropriate choice of $\widetilde{\boldsymbol{W}}$ (i.e. the posterior mode) over 100 simulations for each scenario of Z.

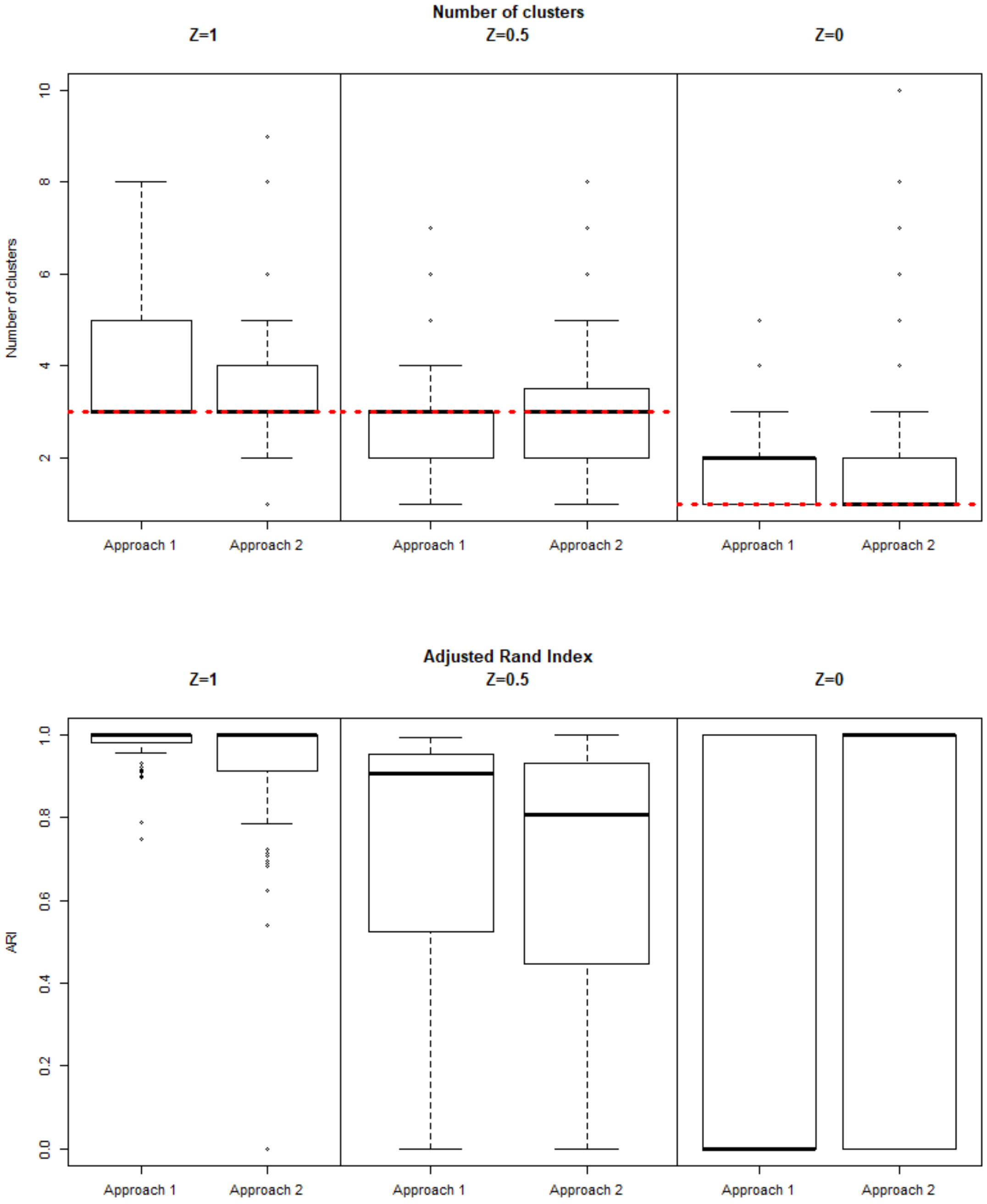


Figure 4.3: A comparison of the results between the proposed modelling **Approach 1** and **Approach 2** in terms of the estimated number of clusters and the adjusted Rand Index (ARI) between the true and estimated cluster structures under each value of Z. In the top panel the dashed lines represent the true number of clusters.

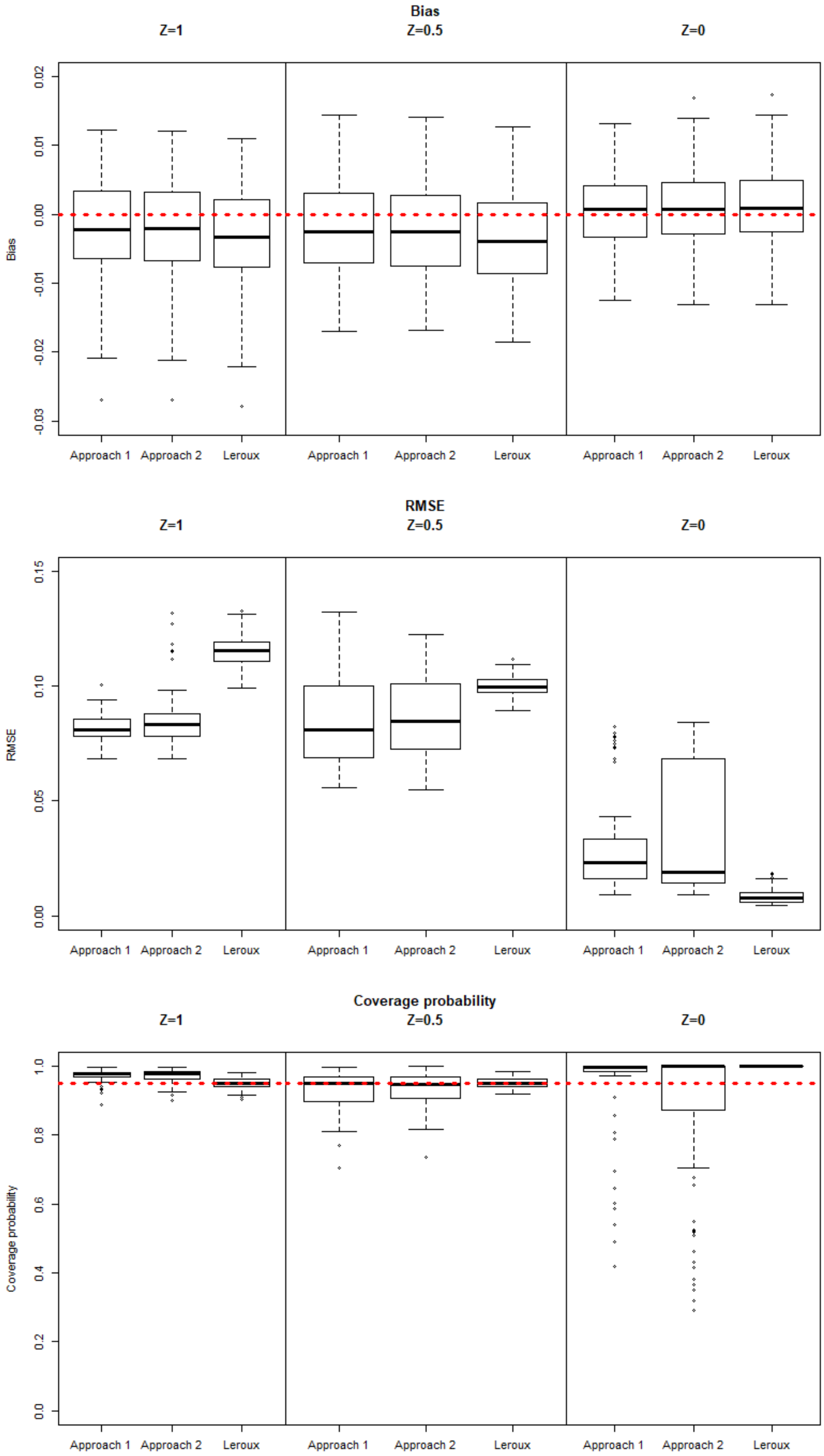


Figure 4.4: A comparison of the proposed modelling approaches and the Leroux model in terms of the bias, root mean square error (RMSE) and 95% coverage probabilities for risk estimates over all simulated data sets. The results relate to $Z = 1$ (left column panels), $Z = 0.5$ (middle column panels) and $Z = 0$ (right column panels). In the bias boxplots the dashed lines represent the zero bias. In the coverage probability boxplots the dashed lines represent the nominal 0.95 levels.

Table 4.3: Summary of the median number of clusters, adjusted Rand Index (ARI), bias, RMSE and 95% coverage probabilities of the estimated risk surface for each model and scenario. Values in brackets display the standard deviation.

		Model		
Performance metric	Z	**Approach 1**	**Approach 2**	**Leroux**
Estimated number of clusters	1	3 (1.189)	3 (1.474)	−−
	0.5	3 (1.222)	3 (1.545)	−−
	0	2 (0.992)	1 (1.877)	−−
ARI	1	1 (0.040)	1 (0.179)	−−
	0.5	0.906 (0.252)	0.806 (0.362)	−−
	0	0 (0.502)	1 (0.492)	−−
Bias	1	-0.002 (0.007)	-0.002 (0.007)	-0.003 (0.007)
	0.5	-0.002 (0.007)	-0.003 (0.007)	-0.004 (0.007)
	0	0.001 (0.005)	0.001 (0.006)	0.001 (0.006)
RMSE	1	0.081 (0.006)	0.083 (0.011)	0.115 (0.006)
	0.5	0.081 (0.018)	0.085 (0.017)	0.100 (0.004)
	0	0.023 (0.019)	0.019 (0.026)	0.008 (0.003)
Coverage probability	1	0.977 (0.016)	0.977 (0.021)	0.951 (0.015)
	0.5	0.949 (0.056)	0.946 (0.047)	0.949 (0.014)
	0	0.996 (0.115)	1 (0.214)	1 (0.000)

4.4 Sensitivity analysis to disease prevalence

In this section, an additional simulation study is carried out to assess the sensitivity of the proposed modelling approaches to the size of the expected disease counts $\boldsymbol{E}$ which quantify the disease prevalence. Modelling **Approach 1** and **Approach 2** are applied to disease data where the expected disease counts $\boldsymbol{E}$ are uniformly drawn from three different intervals: $\boldsymbol{E} \in [10,30]$, $[50,100]$ and $[100,150]$. The simulated data are generated as described in Section 4.3.2. Likewise, three values of $Z = 1, 0.5, 0$ are used in this study. One hundred simulated data sets are generated under each of the nine scenarios comprising pairwise combinations of $Z = 1, 0.5, 0$ and $\boldsymbol{E} \in [10,30], [50,100]$ and $[100,150]$. The results of this analysis are summarised in Figures 4.5 ($Z = 1$), 4.6 ($Z = 0.5$) and 4.7 ($Z = 0$), and outlined in Table 4.4 respectively.

For $Z = 1$, Figure 4.5 shows that the two models estimate the correct number of clusters (3 clusters) on average under $\boldsymbol{E} \in [50,100]$ and $\boldsymbol{E} \in [100,150]$ and identify the correct cluster structure with median ARI values of 1 in both cases. Both modelling **Approach 1** and **Approach 2** perform much better than the Leroux model in terms of lower RMSE values. Under $\boldsymbol{E} \in [50, 100]$ the median RMSE values are 0.089, 0.090 for the proposed approaches and 0.132 for the Leroux, and under $\boldsymbol{E} \in [100, 150]$ the median RMSE values are 0.070 for our approaches and 0.098 for the Leroux. All three models have approximately zero bias of risk estimates and exhibit good coverage probabilities close to their nominal 0.95 levels. Whereas in the case of $\boldsymbol{E} \in [10, 30]$, the proposed models have a high number of simulations either overestimating or underestimating the number of clusters and perform less well in cluster identification with low median ARI values of 0.604 and 0.642 respectively. This is probably because the observed disease counts $\boldsymbol{Y}$ are Poisson distributed with mean equal to $\boldsymbol{E} \times \boldsymbol{R}$, and given risks $\boldsymbol{R}$ fixed, small values of $\boldsymbol{E}$ would make the differences between clusters less prominent in terms of the size of $\boldsymbol{Y}$, which thus makes the cluster identification more difficult and less accurate. Although **Approach 1** and **Approach 2** produce less accurate cluster structures and relatively lower coverage probabilities than the Leroux approach in this case, they still have lower RMSE values (0.228 and 0.223 respectively) compared with the Leroux approach (0.260).

For $Z = 0.5$, the proposed approaches perform very poorly in identifying the correct cluster structures under $\boldsymbol{E} \in [10, 30]$, with median ARI values as low as 0.021 for **Approach 1** and 0.272 for **Approach 2**. Both modelling **Approach 1** and **Approach 2** exhibit slightly higher RMSE values than the Leroux model, with median values of 0.216 and 0.214 respectively compared to 0.203. Besides, the coverage probabilities of our approaches (around 0.86) are much lower than those of the Leroux model (0.95). This is because, given the poor clustering performance, the candidate neighbourhood matrices being fed into the Bayesian spatial model are inherently inaccurate before even being modelled. Thus, the estimation from the proposed approaches is going to be poor and so the coverage is low. Under $\boldsymbol{E} \in [50, 100]$, our approaches generally perform as well as the Leroux approach in terms of both bias and RMSE (see Figure 4.6). However, I note that **Approach 1** and **Approach 2** obtain lower ARI values under $\boldsymbol{E} \in [50, 100]$ with a median around 0.5 than those under $\boldsymbol{E} \in [100, 150]$ with a median close to 1. Again, larger values of $\boldsymbol{E}$ would make the differences between clusters more pronounced in terms of the size of $\boldsymbol{Y}$, which hence makes it much easier to obtain the cluster structures close to the truth. Under $\boldsymbol{E} \in [100, 150]$, all three models perform well with the bias close to zero and the coverage close to the nominal 0.95 levels, while the approaches developed here outperform the Leroux model for having lower RMSE values, with reductions in the median of 44.44% for **Approach 1** and 40% for **Approach 2**.

For $Z = 0$, Figure 4.7 shows that the Leroux model generally performs best of the three in terms of the lowest RMSE values for all scenarios of $\boldsymbol{E}$. This also backs up the findings from Figures 4.3 and 4.4 that the proposed approaches behave slightly less well in estimating a smooth risk surface and identifying the true cluster structure. As explained previously in Section 4.3.3, when the disease risk surface is spatially smooth our approaches are very likely to identify false cluster structures due to the random variation induced into the count data, and the ARI values could be as low as 0. These wrong clusters identified would result in incorrect smoothing of disease risks between neighbouring areal units, hence reducing the estimation accuracy.

The simulation results reported above suggest that the performance of the proposed approaches appears to be affected by the type of disease data to be applied. The two approaches overall perform well and are typically superior to the commonly used Leroux model when the disease is not rare (i.e. the expected disease counts $\boldsymbol{E}$ are not very small).

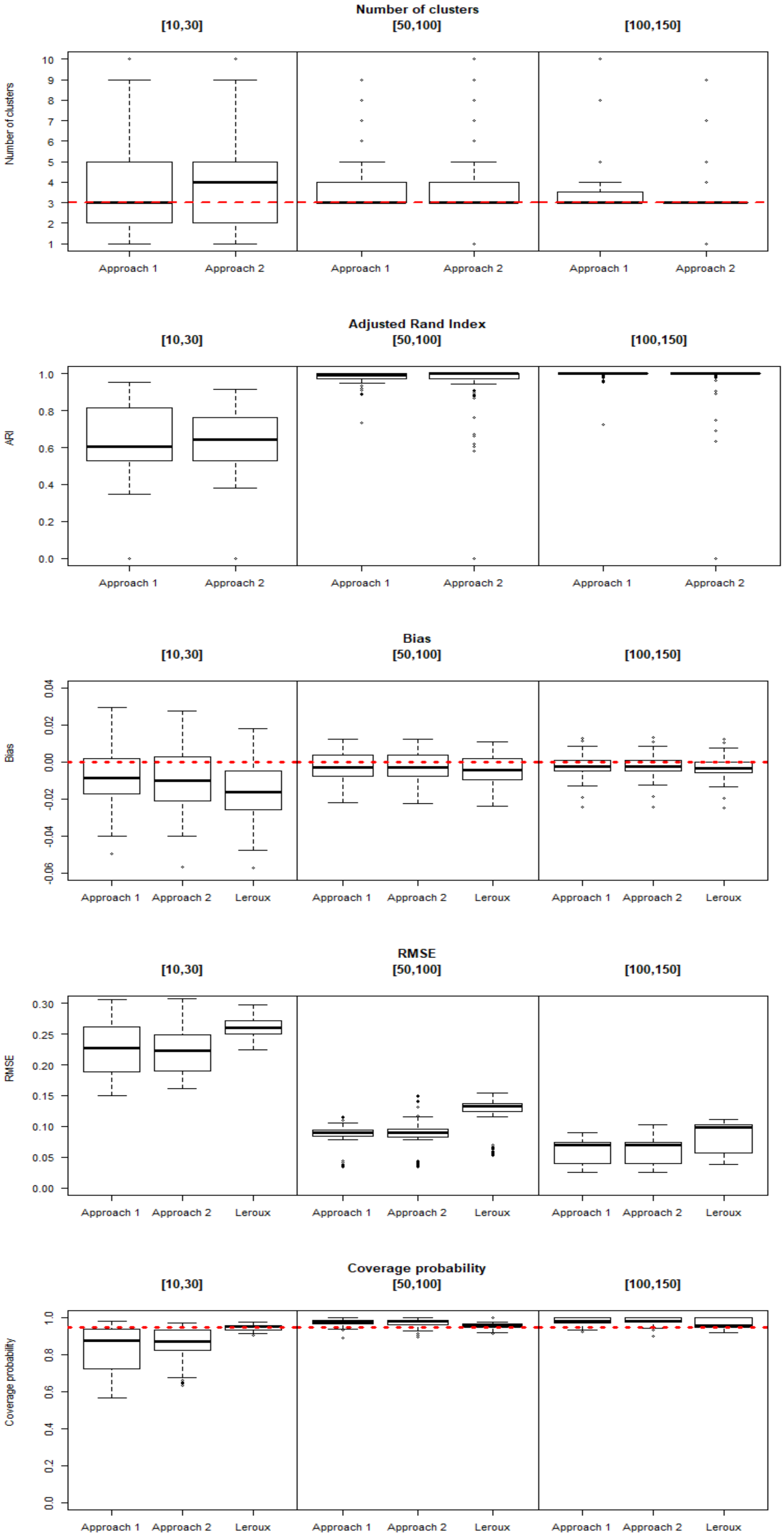


Figure 4.5: Simulation study results from $\boldsymbol{E} \in$ [10,30], [50,100] and [100,150] in terms of the estimated number of clusters, adjusted Rand Index (ARI), bias, RMSE and 95% coverage probability of risk estimates for each model under $Z = 1$.

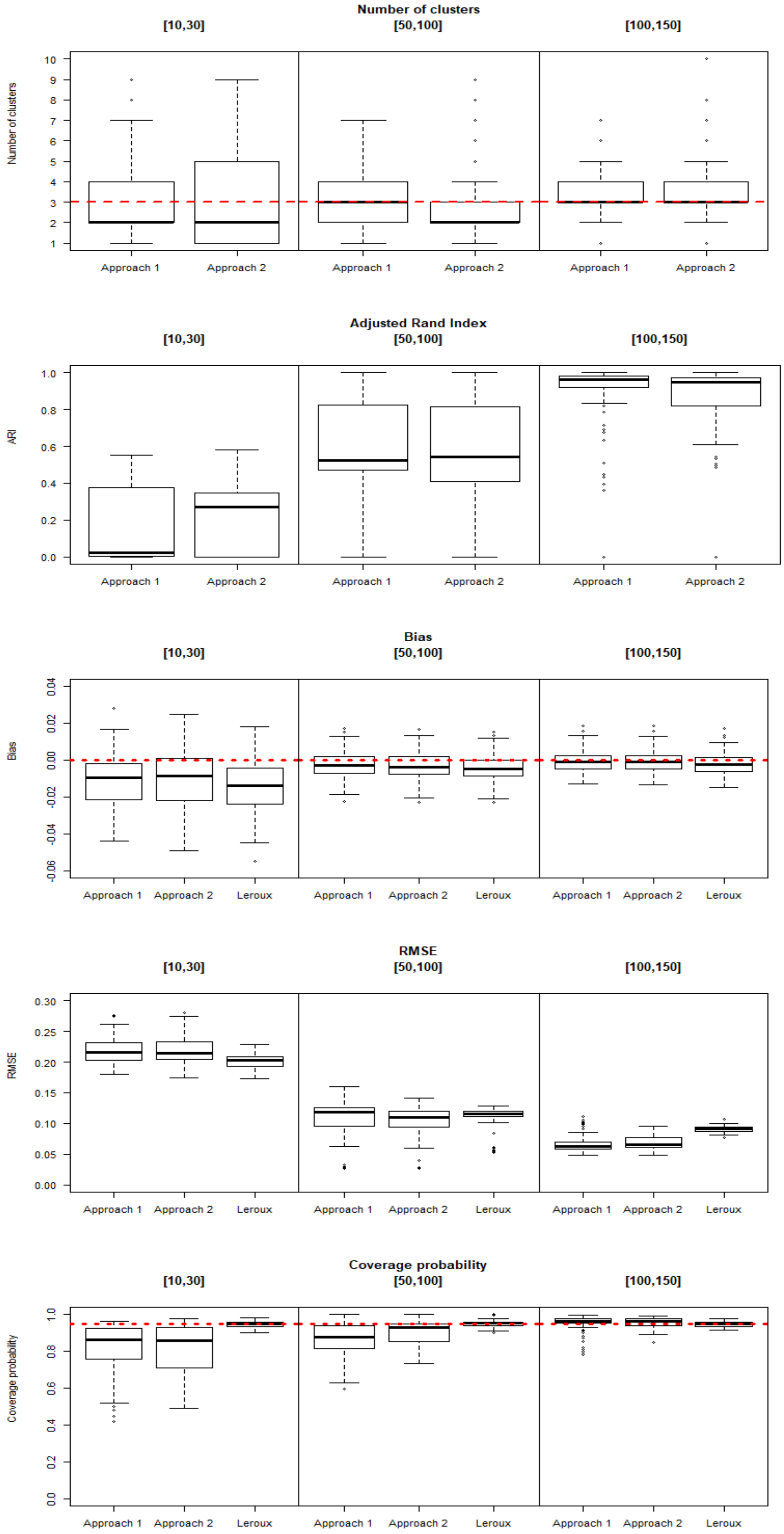


Figure 4.6: Simulation study results from $\boldsymbol{E} \in$ [10,30], [50,100] and [100,150] in terms of the estimated number of clusters, adjusted Rand Index (ARI), bias, RMSE and 95% coverage probability of risk estimates for each model under $Z = 0.5$.

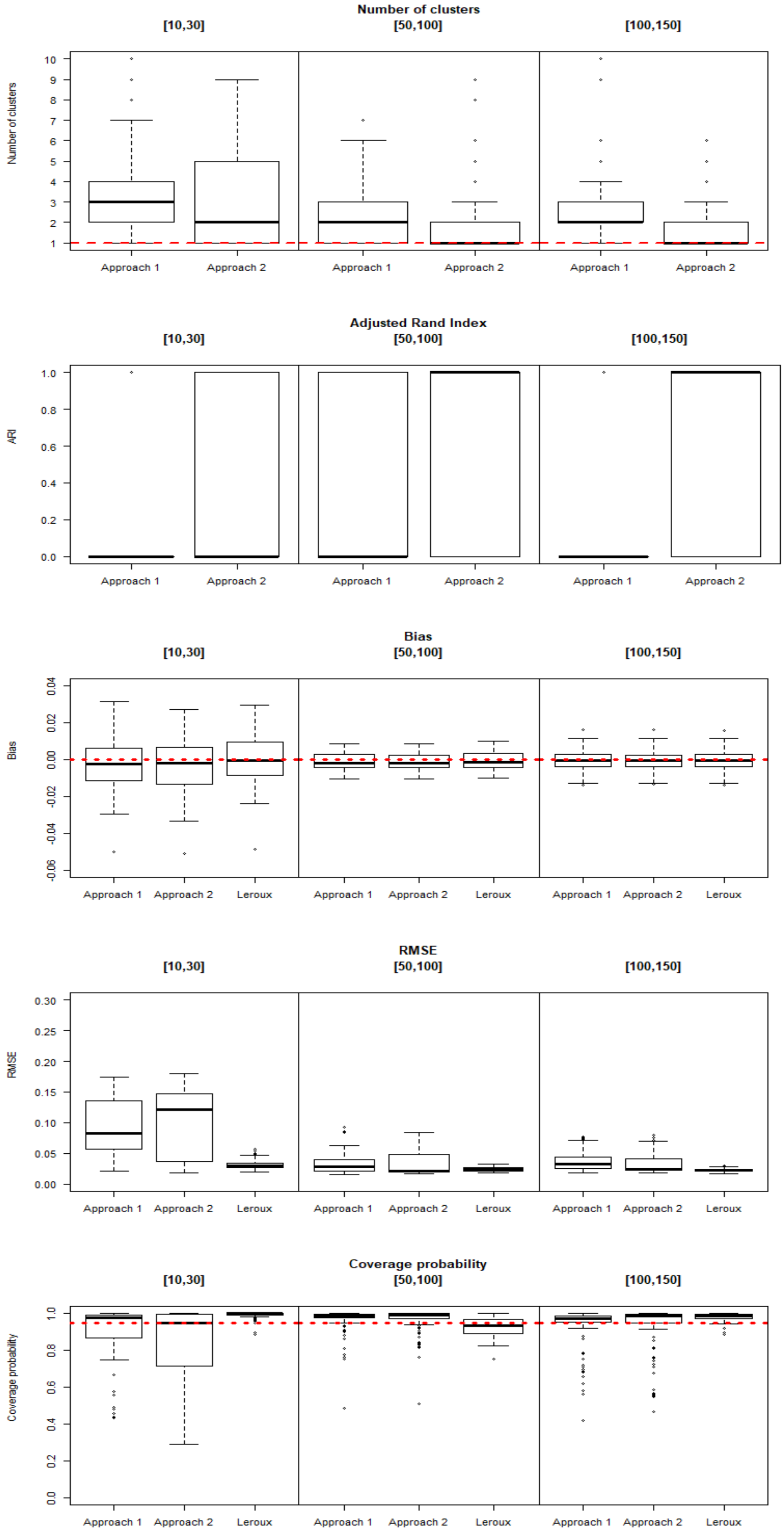


Figure 4.7: Simulation study results from $\boldsymbol{E} \in$ [10,30], [50,100] and [100,150] in terms of the estimated number of clusters, adjusted Rand Index (ARI), bias, RMSE and 95% coverage probability of risk estimates for each model under $Z = 0$.

Table 4.4: Summary of the median number of clusters, adjusted Rand Index (ARI), bias, RMSE and 95% coverage probabilities of the estimated risk surface for each model under each scenario of $\boldsymbol{E}$ and Z. Values in brackets display the standard deviation.

			Model		
Performance metric	$\boldsymbol{E}$	Z	**Approach 1**	**Approach 2**	Leroux
Number of clusters	[10, 30]	1	3 (2.076)	4 (1.994)	−−
		0.5	2 (1.937)	2 (2.290)	−−
		0	3 (2.123)	2 (2.475)	−−
	[50, 100]	1	3 (1.131)	3 (1.487)	−−
		0.5	3 (1.385)	2 (1.654)	−−
		0	2 (1.443)	1 (1.633)	−−
	[100, 150]	1	3 (1.049)	3 (1.228)	−−
		0.5	3 (1.039)	3 (1.736)	−−
		0	2 (1.427)	1 (1.085)	−−
ARI	[10, 30]	1	0.604 (0.167)	0.642 (0.216)	−−
		0.5	0.021(0.198)	0.272 (0.184)	−−
		0	0 (0.402)	0 (0.456)	−−
	[50, 100]	1	0.995 (0.034)	1.000 (0.245)	−−
		0.5	0.524 (0.244)	0.541 (0.329)	−−
		0	0(0.488)	1 (0.476)	−−
	[100, 150]	1	1 (0.029)	1 (0.149)	−−
		0.5	0.964 (0.184)	0.951 (0.298)	−−
		0	0 (0.423)	1 (0.490)	−−
Bias	[10, 30]	1	-0.008 (0.015)	-0.009 (0.016)	-0.016 (0.015)
		0.5	-0.009 (0.015)	-0.008 (0.015)	-0.014 (0.015)
		0	-0.003 (0.014)	-0.002 (0.014)	-0.002 (0.014)
	[50, 100]	1	-0.003 (0.008)	-0.003 (0.008)	-0.004 (0.008)
		0.5	-0.003 (0.008)	-0.004 (0.007)	-0.005 (0.007)
		0	-0.002 (0.005)	-0.002 (0.005)	-0.002 (0.005)
	[100, 150]	1	-0.003 (0.006)	-0.002 (0.006)	-0.003 (0.006)
		0.5	-0.001 (0.006)	-0.001 (0.006)	-0.002 (0.006)
		0	0.000 (0.006)	0.000 (0.006)	0.000 (0.006)
RMSE	[10, 30]	1	0.228 (0.039)	0.223 (0.036)	0.260 (0.016)
		0.5	0.216 (0.021)	0.214 (0.023)	0.203 (0.011)
		0	0.083 (0.047)	0.121 (0.053)	0.029 (0.008)
	[50, 100]	1	0.089 (0.022)	0.090 (0.025)	0.132 (0.029)
		0.5	0.119 (0.030)	0.111 (0.021)	0.116 (0.017)
		0	0.027 (0.016)	0.021 (0.016)	0.023 (0.003)
	[100, 150]	1	0.070 (0.020)	0.070 (0.020)	0.098 (0.027)
		0.5	0.063 (0.014)	0.065 (0.012)	0.091 (0.005)
		0	0.033 (0.016)	0.024 (0.019)	0.022 (0.003)
Coverage probability	[10, 30]	1	0.877 (0.120)	0.875 (0.087)	0.953 (0.015)
		0.5	0.866 (0.151)	0.860 (0.152)	0.949 (0.016)
		0	0.981 (0.148)	0.947 (0.211)	1 (0.020)
	[50, 100]	1	0.981 (0.018)	0.984 (0.022)	0.957 (0.022)
		0.5	0.877 (0.085)	0.930 (0.066)	0.953 (0.020)
		0	0.988 (0.070)	0.994 (0.072)	0.934 (0.048)
	[100, 150]	1	0.984 (0.017)	0.984 (0.017)	0.961 (0.025)
		0.5	0.965 (0.045)	0.965 (0.027)	0.949 (0.014)
		0	0.975 (0.114)	0.988 (0.138)	0.988 (0.021)

4.5 Application to real data

This section continues the analysis of the respiratory hospitalisation data presented in Chapter 3. As displayed in Section 4.3, modelling **Approach 1** slightly outperforms **Approach 2** in terms of estimating more accurate risks and cluster structures, as well as providing less varied results (lower standard variations) in the presence of clusters, therefore **Approach 1** is applied to the respiratory disease data in Greater Glasgow in 2016 which are introduced in Section 3.2, aiming to provide improved risk estimation and identify clusters of areas with different risk levels. The model is run ten times to generate MCMC samples for ten independent Markov chains, and these posterior samples are combined together for overall inference. Each chain is run for 300,000 samples with a burn-in period of 200,000 and thinned by 10. This gives a total of 100,000 samples, with 10,000 samples for each chain. Convergence is checked via visually examining parameter trace plots.

Figure 4.8 presents the estimated spatial risk pattern (posterior median) in Greater Glasgow for 2016. The estimated risks vary between 0.65 and 2.5 over all IZs in the study region with a mean risk of 1.34, suggesting that on average the respiratory disease risk in Greater Glasgow is 34% higher than the overall Scotland in 2016. In addition, it appears that some areas exhibit remarkably different disease risks from their neighbours. For example, the West End area has lower risks than some of its surrounding areas such as Drumchapel and Drumry. Lennoxtown and Milton of Campsie in the north of the city have much lower risks than their neighbour Kirkintilloch. This spatial variation is likely attributed to the socio-economic deprivation, which has been widely evidenced to be linked with disease risks (McCartney, 2012). The high-risk areas in Figure 4.8 generally suffer high levels of socio-economic deprivation as measured by Scottish index of Multiple Deprivation (SIMD), whereas the low-risk areas are wealthier and more prosperous.

In order to illustrate the uncertainty in the estimated cluster structure, Figure 4.9 summarises the posterior samples of $\widetilde{\boldsymbol{W}}$ over 10 Markov chains. In the figure each grid square represents a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k)}$ corresponding to a distinct cluster structure for the data, where the horizontal axis denotes the number of clusters k and the vertical axis denotes the clustering method c. The grid square on the bottom left corner corresponds to the candidate matrices with $k = 1$, representing no clusters in disease risk. The posterior probability of each candidate matrix being identified as the best choice of $\widetilde{\boldsymbol{W}}$ (i.e. the posterior mode) is also provided. It can be seen that the matrices with the top three highest probabilities correspond to the cluster structures generated by k-medoids clustering with $k = 2$ clusters and average-linkage agglomerative clustering with $k = 3$ and $k = 4$ clusters, whose probabilities are 0.291, 0.277 and 0.197 respectively. Figure 4.10 displays these three most likely cluster structures selected by the proposed model, which are denoted by (a), (b) and (c). The blue dots in maps represent the discontinuities detected by the model, which occur between geographical neighbours that are assigned to different clusters due to them exhibiting remarkably different disease risks. As a result, high risks are not smoothed towards their neighbouring areas that have low risks. The cluster structures (b) and (c) show the largest agreement with the highest adjusted Rand Index of 0.880. However, cluster structure (a) is the least similar to the other two structures, with ARI values of 0.492 ((a) vs (b)) and 0.377 ((a) vs (c)) respectively.

The three cluster structures in Figure 4.10 have a number of similarities. They generally identify the same spatially distinct clusters of areas exhibiting high and low risks. The areas of higher risks are predominantly located in the East End of Glasgow containing deprived areas such as Springburn, Easterhouse and Barlanark, and also along the southern bank of the Clyde river including Govan area. The areas of lower risks are mostly less deprived which mainly lie in the affluent West End e.g. Dowanhill, in the far north and north-east of the city e.g. Milngavie and Milton, and to the south e.g. Whitecraigs and Newton Mearns. Cluster structures (a) and (b) have the similar posterior probabilities which are both close to 0.3. The main difference between them is that structure (b) identifies more spatially distinct clusters and discontinuities. In structure (a), all IZs are split into two clusters with a high and low level of disease risk, and the discontinuities occur where the high-risk areas are geographically adjacent to the low-risk areas. However, in cluster structure (b), all IZs are split into three cluster levels containing low, high and very high risk levels. The discontinuities are detected not only where the high/very high risk areas and low risk areas are neighbouring but also where the high risk and very high risk areas are adjacent. For instance, the East End of the city as a whole belongs to the high-risk cluster in structure (a), while in structure (b) some particular areas in the high-risk East End such as Easterhouse and Shettleston are further picked out and identified as being in the very high-risk cluster. In addition, the areas of Inverclyde (e.g. Quarriers, Greenock) in the far west are in a low-risk cluster in cluster structure (a), but they belong to a high-risk cluster in structure (b). Cluster structure (c) has the lowest posterior probability of the three and it partitions the risk surface into four distinct cluster levels. This slightly higher number of clusters would make the average risk levels over these clusters closer to each other, thus some noisy discontinuities are likely to be identified between neighbouring areas that are in different clusters but their risks are not substantially different, e.g. in the north of the map for structure (c). Therefore, cluster structures (a) and (b) are preferable to (c) in this application.

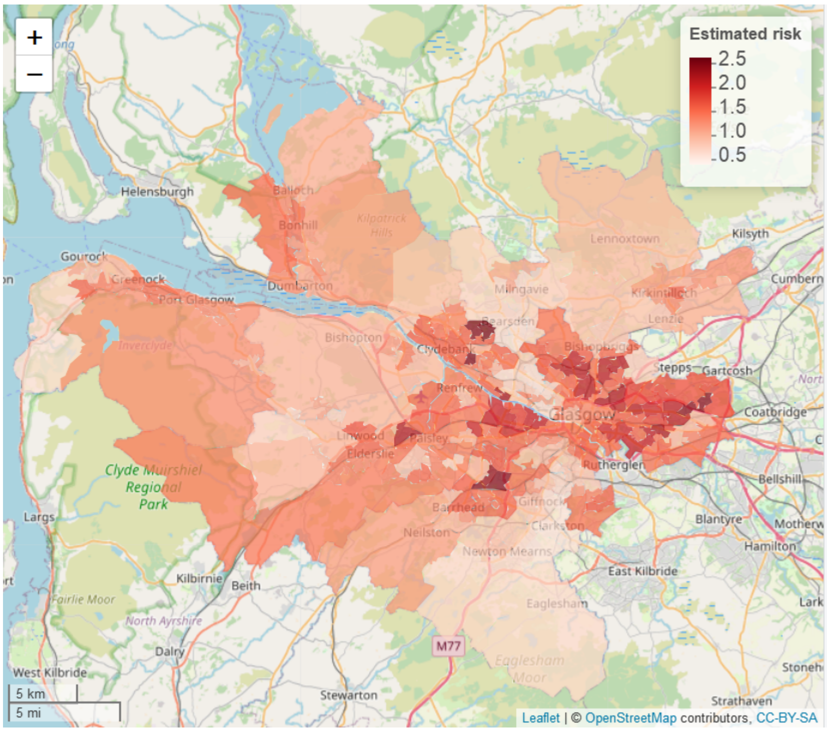


Figure 4.8: Map of the risk estimates (posterior median) for respiratory disease in Greater Glasgow for 2016

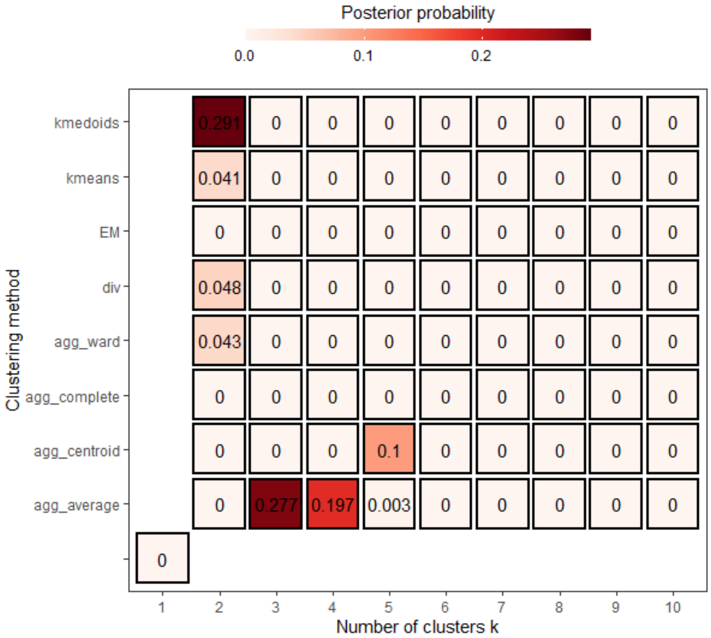
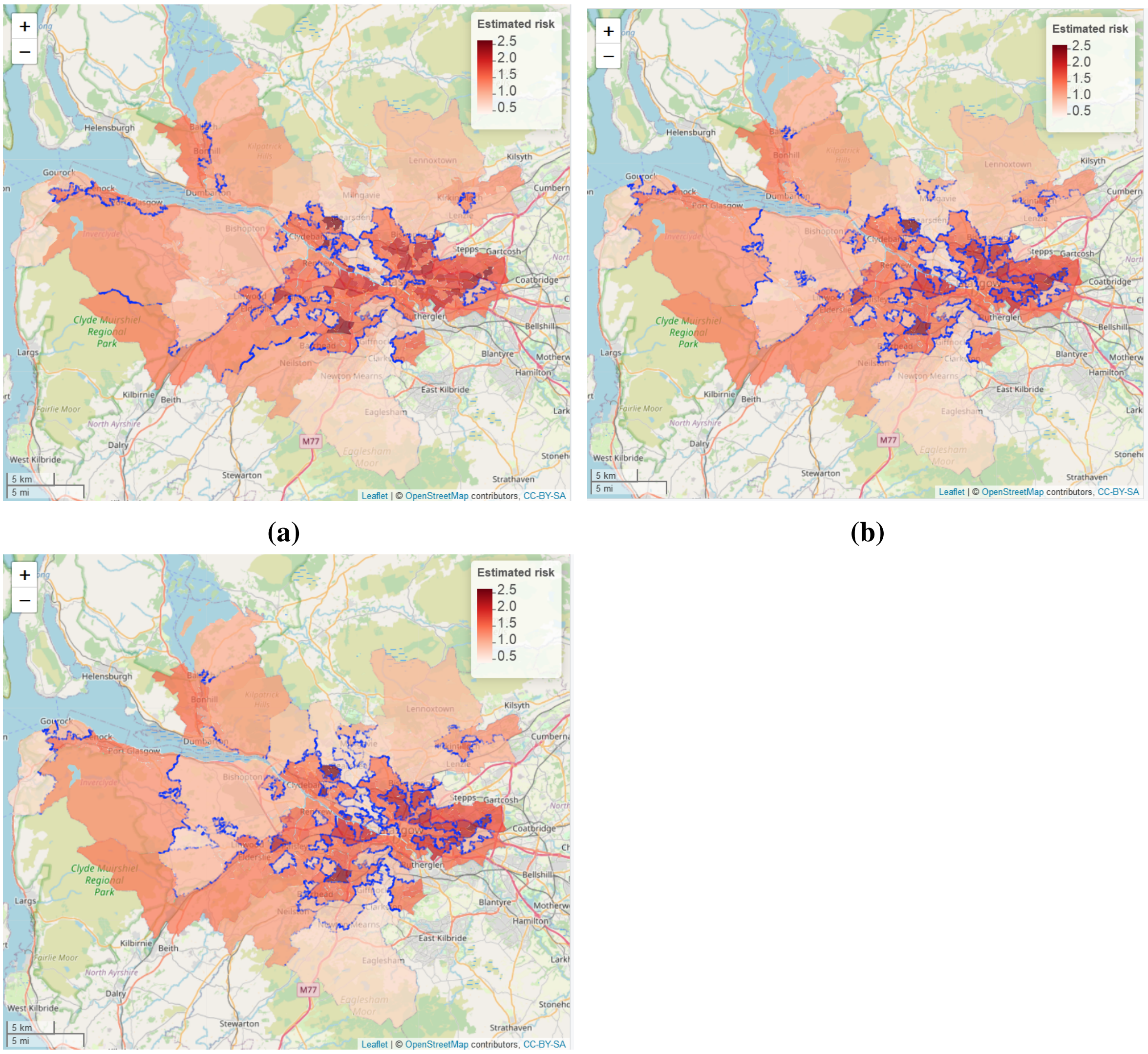


Figure 4.9: Summary of the posterior distribution of $\widetilde{\boldsymbol{W}}$ over ten Markov chains from modelling **Approach 1**



(a)

(b)

(c)

Figure 4.10: The three most likely cluster structures selected by modelling **Approach 1**, with clusters/discontinuities indicated by blue dots. The three structures (a), (b) and (c) respectively identify 2, 3 and 4 cluster risk levels for the data, with the posterior probabilities being 0.291, 0.277 and 0.197. The colour shading for the areas denotes posterior median disease risk in 2016.

4.6 Discussion

In this chapter, I developed the methodology that simultaneously estimates the disease risk surface and identifies clusters of areas with high risks. The basic framework is to elicit a set of cluster structures by partitioning all the areal units into groups based on their SIR values. Each candidate cluster structure is then used to produce a candidate neighbourhood matrix through adapting the border sharing $\boldsymbol{W}$ by setting each element $w_{ij} = 0$ if areas (i, j) are geographically adjacent and in different clusters. The neighbourhood matrix $\widetilde{\boldsymbol{W}}$ is treated as a univariate parameter which is estimated from the set of candidate matrices within the modelling process. Similarly to the method proposed in the previous chapter, the methodology only enforces spatial autocorrelation between neighbouring areas that are in the same cluster, while not allowing for any spatial smoothing of random effects for areas in different clusters. However, unlike the method presented in Chapter 3 where only k-means clustering is used to generate cluster structures in stage one and the neighbourhood matrix is fixed when estimating the other model parameters, the modelling approach introduced here uses a wide range of clustering methods to allow for more flexibility in cluster identification. Furthermore, the uncertainty in $\widetilde{\boldsymbol{W}}$, and hence in the cluster structure identified, is quantified by treating $\widetilde{\boldsymbol{W}}$ as a random parameter whose value is informed mainly by the data. In addition, the methodology has the advantage of reducing the computational time. Rather than fitting multiple Bayesian models separately and then comparing them according to a model selection rule as proposed in Chapter 3, both risk estimation and cluster identification are able to be realised in one single Bayesian spatial model in this chapter.

The simulation studies presented in Section 4.3 and 4.4 show that modelling **Approach 1** and **Approach 2** have advantages over the existing Leroux model fitted based on the border sharing $\boldsymbol{W}$ in certain circumstances. In the presence of discontinuities and clusters in the data, our approaches generally perform well, in particular outperforming the Leroux model in terms of improved performance for estimating risks and identifying clusters of areas with elevated or reduced risks. This is because in contrast to the Leroux model which is set up to estimate a spatially smooth risk surface by always enforcing spatial autocorrelation between geographically neighbours, the proposed methodology accounts for clusters/discontinuities in the spatial autocorrelation structure of the random effects, and attempts to estimate the correlation structure as accurate as possible by ruling out any redundant and incorrect spatial smoothing of the random effects between neighbours. When the disease data are smooth and do not show obviously high or low risk areas, both approaches show slightly poorer accuracy in risk estimation than the Leroux model. This is unsurprising since our approaches would still impose the clustering on the data and expect to find non-smooth patterns from a smooth risk surface. Thus in this case the estimated clusters would be very likely to be incorrect, which further affects the accuracy of risk estimates. Moreover, the simulation studies also suggest that the methodology performs well typically for non-rare diseases with moderate to large number of expected cases. When the expected disease counts are small, e.g. less than 30 cases in each areal unit, the approaches are likely to estimate less accurate risks and cluster structures. Therefore, modelling **Approach 1** and **Approach 2** are more suitable for studying a prevalent or moderately prevalent disease.

Modelling **Approach 1** is applied to the respiratory disease data in Greater Glasgow in 2016 in Section 4.5. The cluster structure with the highest posterior probability partitions the spatial risk pattern into two clusters of areas with a high and low level of disease risk. It identifies the high-risk IZs and differentiates them from the low-risk IZs. The cluster structure with the second highest probability contains three clusters (risk levels), which can also identify the IZs with extremely high risks besides clusters of high- and low-risk areas. One avenue for future work is to extend the spatial methodology proposed here into the spatio-temporal domain. This would help to evaluate the effect of public health policies, and also allow heath authorities to identify clusters of areas that exhibit an increasing disease risk over time. Therefore, in Chapter 5 I will introduce a spatio-temporal model which allows us to estimate how the risk surface and clusters change over time.

Chapter 5

Estimating spatio-temporal disease risks and identifying clusters via clustering-based adjacency modelling

5.1 Introduction

The Bayesian models presented in Chapters 3 and 4 are designed in a purely spatial context to identify the cluster structure and estimate the disease risk for each area at a specific time period. However, since areal unit disease data are typically available for a range of consecutive non-overlapping time periods, investigating the changing nature of disease risk over time has also gained increased popularity. Here I extend the spatial methodology introduced in Chapter 4 to the spatio-temporal domain, with the goal of capturing the spatial pattern of disease risk over time and identifying the, possibly temporally evolving, cluster/discontinuity structures in disease risk. Understanding the temporal trend in disease risk is important because it would allow health authorities to identify groups of areal units where the disease risk has increased over time, and to investigate the potential causes for such deterioration in health. It would also help to design localised disease intervention and evaluate the effect of public health policies. The methodology is motivated by a study of respiratory disease risk in the 257 Intermediate Zones that comprise the Greater Glasgow and Clyde Health Board during the time period from 2011 to 2017.

The remainder of this chapter is organised as follows. Section 5.2 presents the motivating data set, as well as giving a brief review of spatio-temporal disease risk models. Section 5.3 presents the new methodology, while the efficacy and sensitivity of this approach are evidenced using simulations in Section 5.4 and Section 5.5. Section 5.6 applies the methodology to the motivating application, while Section 5.7 provides a discussion on the main findings from the model fitting and ideas for future work.

This chapter is based on the published paper *Spatio-temporal disease risk estimation using clustering-based adjacency modelling*, by Xueqing Yin, Gary Napier, Craig Anderson and Duncan Lee - Statistical Methods in Medical Research (March 14, 2022). DOI: 10.1177/09622802221084131. `https://journals.sagepub.com/doi/10.1177/09622802221084131`.

5.2 Background

5.2.1 Motivating study

The methodology proposed in Section 5.3 is motivated by a study of respiratory disease (defined using the International Classification of Disease tenth revision by codes J00-J99) in the Greater Glasgow and Clyde Health Board region in Scotland between 2011 and 2017, and the study region is shown in Figure 3.1 which consists of $n = 257$ Intermediate Zones (IZs). The disease data, $\boldsymbol{Y} = \{Y_{it}\}$, available from Public Health Scotland, are the yearly counts of the numbers of hospital admissions with a primary diagnosis of respiratory disease for $i = 1, \ldots, n(= 257)$ IZs for $t = 1, \ldots, T(= 7)$ years. Additionally, the expected number of respiratory hospitalisations is calculated for each year and IZ using indirect standardisation to adjust for different population sizes and age and sex structures across the IZs, and are denoted here by $\boldsymbol{E} = \{E_{it}\}$. Specifically, $E_{it} = \sum_{j=1}^{m} n_{itj} r_j$, where n_{itj} is the population size in IZ i, year t and strata j (e.g. females 0-5, females 6-10, etc), and r_j is the Scotland-wide disease rate in strata j. These expected counts are based on Scotland-wide age-sex specific respiratory hospitalisation rates, because it allows us to examine how the disease risk in Glasgow compares to the national average, which is the benchmark often used by Public Health Scotland when examining the spatial patterns in disease risk.

The standardised incidence ratio, $\text{SIR}_{it} = \frac{Y_{it}}{E_{it}}$ is an exploratory (noisy) estimate of disease risk, where a value greater than 1 corresponds to an increased level of risk compared to the Scottish average while a value less than 1 indicates a decreased level of risk than the Scottish average. The spatial patterns in the SIR for 2011 (the first year of data) and 2017 (the last year of data) are displayed in the top and middle panels of Figure 5.1.

They show that higher respiratory disease risks are mainly in the East End of Glasgow (the east of the map) and along the southern bank of the River Clyde, which includes the socio-economically deprived areas of Easterhouse and Govan respectively, both of which are well known to suffer from multi-generational poverty (Glasgow City Council, 2020). In addition, there are numerous pairs of neighbouring areas where a discontinuity in disease risk appears to exist, suggesting the presence of clusters of areas that exhibit elevated risks compared with their neighbours. For example, in 2017 Drumchapel and Drumry to the north west of the city exhibits a vastly higher SIR value (SIR = 1.85) than its neighbour Bearsden (SIR = 0.53). Therefore, the common approach in the literature of assuming that all pairs of neighbouring areal units exhibit similar disease risks is clearly not appropriate, which motivates the spatio-temporal clustering model proposed below.

The spatial patterns in the SIR are fairly similar for each year, with an average Pearson's correlation coefficient of 0.844 between each pair of years. This suggests that while any clusters identified in disease risk may evolve slightly over time, one would not expect a large change in the clusters from year to year. This gradual change is likely to be because respiratory related hospitalisations are a marker of chronic rather than epidemic disease, and hence any change would likely be gradual and due to factors such as the gentrification of an area. The temporal trend in the SIR is shown in the bottom panel of Figure 5.1, which shows that overall there has been a slight increase in the SIR over the 7 year period, with a mean value of 1.10 in 2011 compared to 1.28 in 2017. There also appears to be increased spatial variation in risk in the later years, with standard deviations over space of 0.39 in 2011 and 0.46 in 2017. Finally I do not collect any covariate data for this study, because the aim is to estimate the spatio-temporal trend in disease risk and its spatial cluster structure, rather than the drivers of elevated disease risks. Furthermore, as the clusters identified by the methodology are in the random effects surface, including covariates in the model would mean that the clustering/discontinuities relate to the residual risk surface after covariate adjustment. In contrast, by not including covariates in the model the random effects and risk surfaces have the same spatial structure (see Section 5.3), and thus any clusters/discontinuities identified also relate to disease risk.

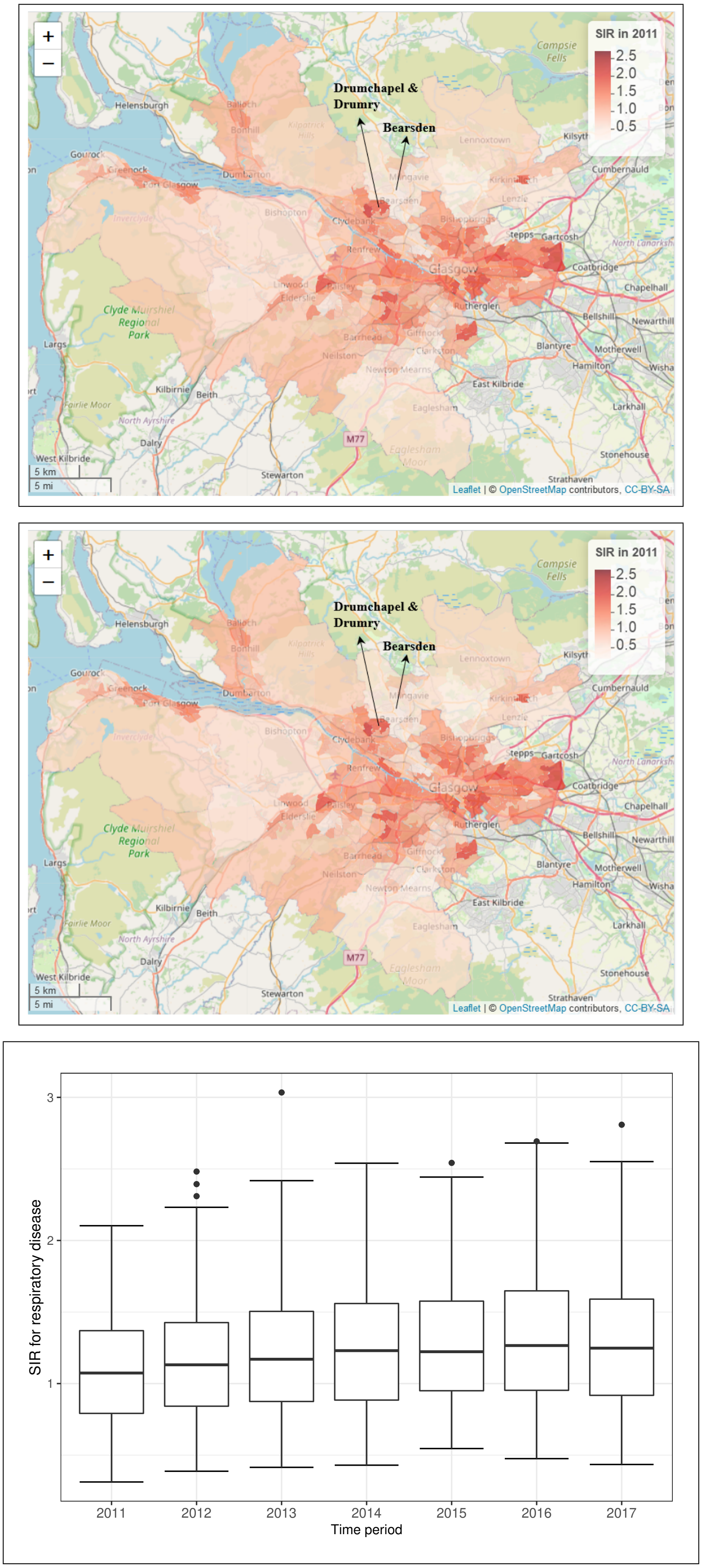


Figure 5.1: Maps of the SIR for respiratory disease in the Intermediate Zones in the Greater Glasgow and Clyde Health Board in 2011 (the top panel) and 2017 (the middle panel). The bottom panel shows boxplots of the SIR over time.

5.2.2 Review of spatio-temporal modelling of areal unit count data

5.2.2.1 Notation and data likelihood

Consider a study region partitioned into n non-overlapping areal units indexed by $i \in \{1,\ldots,n\}$, where data are collected for $t \in \{1,\ldots,T\}$ consecutive time periods. As previously described, $\boldsymbol{Y} = \{Y_{it}\}$ and $\boldsymbol{E} = \{E_{it}\}$ denote the set of observed and expected disease counts respectively, while a vector of covariates (if needed) is given by $\boldsymbol{x}_{it}$ for area i and time period t. The model outlined in this chapter is the most general form that includes covariates, but as previously mentioned the application only includes an intercept term in the model. As the response variable is a count the data likelihood model commonly used is given by $Y_{it}|E_{it},R_{it} \sim \text{Poisson}(E_{it}R_{it})$, where R_{it} represents disease risk in areal unit i during time period t and is on the same scale as the SIR.

5.2.2.2 Spatio-temporal risk model

The spatio-temporal structure in risk $\{R_{it}\}$ is typically modelled by both covariates and random effects, and a large number of different random effects structures have been proposed. An appropriate choice of structure depends on both the aims of the analysis and the trends observed in the data, and the first paper in this area proposed using spatially correlated linear time trends (Bernardinelli et al., 1995) for each area. Probably the most widely used structure to date was proposed by Knorr-Held (2000), which decomposes disease risk into separate spatial and temporal main effects and an additional spatio-temporal interaction term. More recently, two popular spatio-temporal structures were proposed by Rushworth et al. (2014) and Napier et al. (2016) respectively. Rushworth et al. (2014) modelled the risk surface as a spatially autocorrelated multivariate first order autoregressive process, while Napier et al. (2016) built on the model of Waller et al. (1997) by using a region-wide temporal trend and separate spatial processes for each year. In this chapter the latter of these is used, because it is the only one of the aforementioned approaches to have a separate spatial process with a potentially different neighbourhood matrix $\boldsymbol{W}$ for each time period. Having a separate $\boldsymbol{W}$ for each time period is crucial for the methodology proposed in the next section, because it is the mechanism by which we estimate the temporally evolving clusters/discontinuities in disease risk. The risk model proposed by Napier et al. (2016) that our approach is based on is given by the Poisson log-linear specification

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}), \quad i=1,\ldots,n; \quad t=1,\ldots,T, \\ \ln(R_{it}) &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t, \end{aligned} \tag{5.1}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters. The residual spatio-temporal structure (after covariate adjustment) is modelled by an overall temporal trend $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_T)$ and a separate spatial surface at each time period t, $\boldsymbol{\phi}_t = (\phi_{1t},\ldots,\phi_{nt})$. Each spatial surface $\boldsymbol{\phi}_t$ is modelled by the Leroux CAR prior (Leroux et al., 2000) which induces spatial autocorrelation into the random effects via a neighbourhood matrix $\boldsymbol{W}$ that determines which pairs of areal units are close together. Here the commonly used *sharing a common border specification* is adopted, where $w_{ij}=1$ if areal units (i,j) share a common geographical border, and $w_{ij}=0$ otherwise. Based on this matrix the Leroux CAR prior (Leroux et al., 2000) for $\boldsymbol{\phi}_t$ is specified by n univariate full conditional distributions, which for area i is given by

$$\phi_{it}|\boldsymbol{\phi}_{-i,t},\boldsymbol{W} \sim \text{N}\left(\frac{\rho_s\sum_{j=1}^{n}w_{ij}\phi_{jt}}{\rho_s\sum_{j=1}^{n}w_{ij}+1-\rho_s}, \frac{\tau_t^2}{\rho_s\sum_{j=1}^{n}w_{ij}+1-\rho_s}\right), \tag{5.2}$$

where $\boldsymbol{\phi}_{-i,t} = (\phi_{1,t},\ldots,\phi_{i-1,t},\phi_{i+1,t},\ldots,\phi_{n,t})$. The strength of the spatial autocorrelation is controlled by a temporally invariant parameter ρ_s, where a value of 1 indicates strong dependence in space (corresponding to the intrinsic CAR model (Besag et al., 1991)) and a value of 0 indicates spatial independence (as $\phi_{it} \sim \text{N}(0,\tau_t^2)$). Additionally, τ_t^2 is a temporally-varying variance parameter, thus allowing the amount of spatial variation in the data to change over time. The joint multivariate Gaussian distribution for $\boldsymbol{\phi}_t$ corresponding to the above is $\boldsymbol{\phi}_t \sim \text{N}\left(\boldsymbol{0}, \tau_t^2\boldsymbol{Q}(\rho_s,\boldsymbol{W})^{-1}\right)$, where $\boldsymbol{Q}(\rho_s,\boldsymbol{W}) = \rho_s(\text{diag}(\boldsymbol{W}\boldsymbol{1})-\boldsymbol{W})+(1-\rho_s)\boldsymbol{I}$, $\boldsymbol{1}$ is an $n\times 1$ vector of ones and $\boldsymbol{I}$ is an $n\times n$ identity matrix. The partial correlation between (ϕ_{it},ϕ_{jt}) conditioning on the remaining spatial random effects (denoted $\boldsymbol{\phi}_{-ijt}$) specified by this model is

$$\text{Corr}\left(\phi_{it},\phi_{jt}|\boldsymbol{\phi}_{-ijt}\right) = \frac{\rho_s w_{ij}}{\sqrt{\left(\rho_s\sum_{v=1}^{n}w_{iv}+1-\rho_s\right)\left(\rho_s\sum_{v=1}^{n}w_{jv}+1-\rho_s\right)}}. \tag{5.3}$$

Equation (5.3) shows that (ϕ_{it},ϕ_{jt}) are modelled as partially correlated if $w_{ij}=1$, otherwise, the partial correlation between (ϕ_{it},ϕ_{jt}) is 0 and they are modelled as conditionally independent. Hence the neighbourhood matrix $\boldsymbol{W}$ determines the spatial autocorrelation structure imposed by the model. Thus using the border sharing rule the spatial random effect ϕ_{it} is forced to be correlated with its geographical neighbours if ρ_s is estimated as close to one, meaning that any discontinuities in the spatial surface are smoothed over in the estimation. This not only leads to poorer risk estimation in the presence of discontinuities as shown in Section 5.4, but also does not allow a mechanism for identifying these discontinuities that correspond to cluster boundaries. Finally, Napier et al. (2016) modelled the temporal trend $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ by a one dimensional Leroux CAR prior, and further details can be found in Section 2.5.

5.3 Methodology

This chapter proposes a two-stage modelling approach for spatio-temporal clustering that jointly estimates the spatio-temporal pattern in disease risk and identifies clusters of areas with elevated or reduced risks compared to their geographical neighbours. In stage one a range of clustering methods are used to identify a large collection of plausible candidate cluster structures for the data, each of which is then used to create a candidate neighbourhood matrix. In stage two, a Bayesian hierarchical model is proposed for the data, which jointly estimates the spatio-temporal pattern in disease risk and the most appropriate neighbourhood matrix corresponding to a given cluster structure. I propose two different variants of the model, with variant A having spatial clusters that remain fixed during the entire study period, while in variant B the clusters vary dynamically over time.

5.3.1 Stage 1 — Generating neighbourhood matrices representing clusters/discontinuities

As in Chapter 4, a collection of candidate cluster structures for disease risk are estimated based on $c = 1, \ldots, M$ different clustering methods, and the $M = 8$ methods considered here are summarised in Table 4.1. Again, the cluster methods include k-means (MacQueen et al., 1967) clustering, k-medoids (Park and Jun, 2009) clustering, hierarchical agglomerative (Hastie et al., 2009) clustering with centroid, complete, average and Ward linkage, divisive (Kaufman and Rousseeuw, 2009) clustering and expectation-maximisation (Fraley and Raftery, 2002) clustering. These clustering methods are applied to the data without regard to the spatial positions of the areal units, because the spatial correlation in the data is modelled by random effects as described above. Thus the clusters identified represent the number of different risk levels and not the number of spatially distinct clusters, meaning that a single *"cluster"* will likely contain groups of areas that are not spatially connected by $\boldsymbol{W}$.

Each clustering method c is used to compute $k = 1,\ldots,K$ distinct cluster structures, where structure k contains k clusters. As in the previous chapters, the value of K is the upper limit for the number of clusters that is expected to find in the data, and it has to be specified by the user. Here I set $K = 10$ again as a conservative overly large choice, because as described above this represents the number of distinct risk levels and not the number of spatially contiguous clusters. These candidate cluster structures are incorporated into the disease risk model by specifying a set of candidate neighbourhood matrices, which means they relate to the random effects surface $\{\phi_{it}\}$. Therefore the first step to estimating an appropriate neighbourhood matrix is to estimate $\{\phi_{it}\}$ from the data and the general model (5.1) by

$$\tilde{\phi}_{it} \;=\; \ln\left(\frac{\mathbb{E}(Y_{it})}{E_{it}}\right) - \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} - \theta_t \approx \ln\left(\frac{Y_{it}}{E_{it}}\right) - \boldsymbol{x}_{it}^{\top}\hat{\boldsymbol{\beta}}. \tag{5.4}$$

This approximation replaces the unknown $\mathbb{E}(Y_{it})$ with the observed data Y_{it}, and the temporal random effects are removed as they do not vary over space and hence do not affect the spatial cluster structure. The regression parameters are estimated for this initial stage assuming independence via maximum likelihood estimation, and are denoted above by $\hat{\boldsymbol{\beta}}$. These estimated spatial random effects $\{\tilde{\phi}_{it}\}$ are used to construct candidate cluster structures and corresponding neighbourhood matrices as described below for variants A (static) and B (dynamic) of our model.

5.3.1.1 Variant A: Constant cluster structure over time

Temporally constant clusters are obtained by applying each clustering method to $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\phi}}_1,\ldots,\tilde{\boldsymbol{\phi}}_n)$, where $\tilde{\boldsymbol{\phi}}_i = (\tilde{\phi}_{i1},\ldots,\tilde{\phi}_{iT})$. The cluster structure obtained from method c with k clusters is used to create a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k)}$ as follows:

$$w_{ij}^{(c,k)} = \begin{cases} 1, & \text{if areal units } (i,j) \text{ share a common border and are in the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

Thus there is a one-to-one relationship between a candidate cluster structure and its corresponding neighbourhood matrix, and as trivially the border sharing matrix equals all of $\{\boldsymbol{W}^{(1,1)},\ldots,\boldsymbol{W}^{(M,1)}\}$ and this leads to $(K-1)\times M+1$ candidate cluster structures in total. Altering the border sharing $\boldsymbol{W}$ in this way to allow for clusters means that if areas (i,j) share a border and are in the same cluster (i.e. have similar data values) then their random effects will be modelled as partially correlated (see equation (5.3)). In contrast, if they share a border but are in different clusters (i.e. have very different data values) then their random effects will be modelled as conditionally independent, thus not enforcing spatial smoothing between them.

5.3.1.2 Variant B: Temporally varying cluster structures

In variant A the spatial clusters do not change over time, which allows one to estimate an overall average cluster structure in the data across the entire study period. However this may not be realistic in practice, because different areas can have different temporal trends in disease risk leading to evolution in the spatial cluster structure over time. To account for this I adjust the clustering method in variant A by estimating a separate spatial cluster structure for each time period. This is achieved by applying clustering method c with k clusters to $\tilde{\boldsymbol{\phi}}_t = (\tilde{\phi}_{1t}, \ldots, \tilde{\phi}_{nt})$ separately for each time period t, yielding a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k,t)}$ defined as follows:

$$w_{ij}^{(c,k,t)} = \begin{cases} 1, & \text{if areal units } (i,j) \text{ share a common border and are in the same cluster at time } t, \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

This algorithm leads to $(K-1) \times M \times T + 1$ candidate cluster structures in total, including $(K-1) \times M$ for each time period t and the additional border sharing specification.

5.3.2 Stage 2 — Bayesian spatio-temporal modelling

The second stage of the methodology fits a model to the data that jointly estimates the spatio-temporal trend in disease risk and an appropriate cluster/discontinuity structure(s), the latter being achieved by treating the neighbourhood matrix as a parameter to be estimated from the set of candidates generated in stage 1. The model structure proposed here is based on Napier et al. (2016) described in Section 5.2.2, because its separate random effects surfaces for each time period allow for different neighbourhood matrices to be specified in each case, which is necessary under variant B of the model. The first level of the model is given by

$$\begin{aligned} Y_{it}|E_{it}, R_{it} &\sim \text{Poisson}(E_{it}R_{it}) \;\; i = 1, \ldots, n; \;\; t = 1, \ldots, T, \\ \ln(R_{it}) &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t, \\ \beta_j &\sim \text{N}(0, 1000), \text{ for } j = 0, \ldots, p. \end{aligned} \tag{5.7}$$

The residual spatio-temporal variation in the data is decomposed into an overall temporal trend common to all areal units and separate spatial surfaces for each time period. The former is denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ and modelled by the first order autoregressive process

$$\begin{aligned}
\theta_t | \theta_{t-1} &\sim \mathrm{N}\left(\alpha \theta_{t-1}, \sigma^2\right) \text{ for } t = 2, \ldots, T, \\
\theta_1 &\sim \mathrm{N}\left(0, \sigma^2\right), \\
\alpha &\sim \text{Uniform}(0,1), \\
\sigma^2 &\sim \text{Inverse-Gamma}(1, 0.01).
\end{aligned} \tag{5.8}$$

Here $\alpha \in [0,1]$ is the temporal autoregressive parameter, with $\alpha = 1$ indicating strong temporal dependence (a first order random walk), while $\alpha = 0$ corresponds to independence across time. A uniform prior on the interval $[0,1]$ is assigned to α, while a conjugate Inverse-Gamma prior is assigned to the process variance σ^2. The spatial surface at time period t is captured by $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{nt})$, which is modelled by a separate Leroux CAR prior (Leroux et al., 2000) for each time period. Different spatial variances for each time period are allowed because the exploratory analysis in Section 5.2.1 suggested that the level of spatial variation may change over time. For both model variants A and B the neighbourhood matrix is treated as a parameter to be estimated, and the model specifications are given below.

5.3.2.1 Variant A: Constant cluster structure over time

In this model variant there is a single neighbourhood matrix $\widetilde{\boldsymbol{W}}$ that is common to all time periods, and the spatial random effects $\boldsymbol{\phi}_t$ are modelled by:

$$\begin{aligned}
&\boldsymbol{\phi}_t \sim \mathrm{N}\left(\mathbf{0}, \tau_t^2 \boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}})^{-1}\right), \\
&\widetilde{\boldsymbol{W}} \sim \text{Discrete uniform}\left(\boldsymbol{W}^{(1,1)}, \boldsymbol{W}^{(1,2)}, \ldots, \boldsymbol{W}^{(1,K)}, \boldsymbol{W}^{(2,2)}, \ldots, \boldsymbol{W}^{(2,K)}, \ldots, \boldsymbol{W}^{(M,2)}, \ldots, \boldsymbol{W}^{(M,K)}\right).
\end{aligned}$$

Here $\boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}})$ is the spatial precision matrix corresponding to the Leroux CAR prior, which is defined in the previous section. This matrix depends on the neighbourhood matrix $\widetilde{\boldsymbol{W}}$, which is assigned a discrete uniform prior whose possible values are the set of candidates corresponding to the cluster structures estimated in stage 1. Finally, the variance and spatial dependence parameters are assigned weakly informative Inverse-Gamma ($\tau_t^2 \sim$ Inverse-Gamma$(1, 0.01)$) and uniform ($\rho_s \sim$ Uniform$(0,1)$) priors respectively.

5.3.2.2 Variant B: Temporally varying cluster structures

In this model variant there is a different neighbourhood matrix $\widetilde{\boldsymbol{W}}_t$ for each time period t, and the spatial random effects $\boldsymbol{\phi}_t$ are modelled by:

$$\begin{aligned}
\boldsymbol{\phi}_t &\sim \mathrm{N}\left(\mathbf{0}, \tau_t^2 \boldsymbol{Q}(\rho_{s_t}, \widetilde{\boldsymbol{W}}_t)^{-1}\right), \\
\widetilde{\boldsymbol{W}}_t &\sim \text{Discrete uniform}\left(\boldsymbol{W}^{(1,1,t)}, \boldsymbol{W}^{(1,2,t)}, \ldots, \boldsymbol{W}^{(1,K,t)},\right. \\
&\qquad \left.\boldsymbol{W}^{(2,2,t)}, \ldots, \boldsymbol{W}^{(2,K,t)}, \ldots, \boldsymbol{W}^{(M,2,t)}, \ldots, \boldsymbol{W}^{(M,K,t)}\right).
\end{aligned} \tag{5.9}$$

Here $\boldsymbol{Q}(\rho_{s_t}, \widetilde{\boldsymbol{W}}_t)$ again corresponds to the Leroux CAR prior, where in this model variant the spatial dependence parameter ρ_{s_t} changes over time as the neighbourhood matrix also varies over time. As before the set of candidate neighbourhood matrices at time t that make up the discrete uniform prior for $\widetilde{\boldsymbol{W}}_t$ are obtained from the candidate cluster structures generated in stage 1. In common with variant A the model specification is completed with $\tau_t^2 \sim \text{Inverse-Gamma}(1, 0.01)$ and $\rho_{s_t} \sim \text{Uniform}(0, 1)$.

For both model variants the clustering stage elicits multiple candidate neighbourhood matrices that are equal to the border sharing $\boldsymbol{W}$, which occur when the number of clusters $k = 1$ for each clustering method. Therefore only one of these is included in the discrete uniform prior to avoid the border sharing $\boldsymbol{W}$ being given a larger prior weight compared to the other candidate values. This border sharing $\boldsymbol{W}$ is included in the model because it corresponds to a globally spatially smooth risk surface with no clusters. Additionally, to achieve identifiability, all sets of spatial and temporal random effects are zero-mean centred.

Here (ρ_s, ρ_{s_t}) control the level of spatial autocorrelation globally across the study region, with values close to 1 corresponding to strong spatial autocorrelation while a value of zero corresponds to spatial independence. However, in the two model variants the spatial autocorrelation structure is modelled locally for each pair of neighbouring areas by estimating an appropriate neighbourhood matrix for the data, which may make the estimation of a single global parameter redundant. Thus in the simulation study in Section 5.4 I compare the performance of both model variants when estimating (ρ_s, ρ_{s_t}) in the model and also when fixing them at $\rho_s, \rho_{s_t} = 0.99$. The latter is chosen because it is close to one and hence enforces strong spatial autocorrelation globally, whose structure is then adjusted locally by estimating the neighbourhood matrix. Note, $\rho_s, \rho_{s_t} = 1$ is not used because our model could produce a candidate neighbourhood matrix where an areal unit has no neighbours due to it being a singleton cluster. This will cause $\sum_{j=1}^{n} w_{ij} = 0$ for the area i in question, which leads to an infinite mean and variance for ϕ_{it} from (5.2).

5.3.3 Inference

Inference is carried out in a Bayesian setting via Markov chain Monte Carlo (MCMC) simulation, using both the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970) and Gibbs sampling (Geman and Geman, 1984). The only non-standard step is the updating of the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ or $\widetilde{\boldsymbol{W}}_t$, which is achieved by using a Metropolis-Hastings step consisting of two MCMC moves. This is initially introduced in Chapter 4 and briefly outlined below. Note, the step is outlined for variant A of the model, and the updating step for variant B is analogous.

1. If the current value of $\widetilde{\boldsymbol{W}}$ in the Markov chain is $\boldsymbol{W}^{(c,k)}$, then a new value $\boldsymbol{W}^{(c,l)}$ is proposed uniformly from the set of candidate matrices $\left(\boldsymbol{W}^{(c,k-s)},\ldots,\boldsymbol{W}^{(c,k-1)},\boldsymbol{W}^{(c,k+1)},\ldots,\boldsymbol{W}^{(c,k+s)}\right)$, which is the candidate cluster structures generated from the same clustering method but with a different number of clusters. Here s is a parameter controlling the acceptance rates and mixing of the update, and exploratory model runs suggested that $s = 2$ leads to good estimation performance.

2. If the current value of $\widetilde{\boldsymbol{W}}$ after the first move is $\boldsymbol{W}^{(c,k')}$, then a new proposal $\boldsymbol{W}^{(h,k')}$ is drawn uniformly from the set $\left(\boldsymbol{W}^{(1,k')},\ldots,\boldsymbol{W}^{(c-1,k')},\boldsymbol{W}^{(c+1,k')},\ldots,\boldsymbol{W}^{(M,k')}\right)$, which is the candidate cluster structures with the same number of clusters but generated from a different clustering method.

Since the neighbourhood matrix follows a discrete distribution, the posterior mode of $(\widetilde{\boldsymbol{W}},\ \widetilde{\boldsymbol{W}}_t)$, representing the most likely occurring cluster structure across all the MCMC samples, is used to estimate the optimal cluster/discontinuity structure. In contrast, the remaining parameters are summarised by their posterior medians. The MCMC algorithm for fitting the model was developed and implemented in R (R Core Team, 2013) and C++ via the R package Rcpp (Eddelbuettel et al., 2011) and is available from `https://github.com/XueqingYin/ST-model`. Details of each step of the MCMC sampler for the model are given as follows.

Update $\boldsymbol{\beta}$

The full conditional distribution for $\boldsymbol{\beta}$ is

$$\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{Y}) &\propto \prod_{i=1}^{n}\prod_{t=1}^{T}\text{Poisson}(Y_{it}|\boldsymbol{\beta}) \times \prod_{j=0}^{p}\text{N}(\beta_j|0,1000) \\
&\propto \left(\prod_{i=1}^{n}\prod_{t=1}^{T}\left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it}+\theta_t)\right)^{Y_{it}}\right)\exp\left(-\sum_{i=1}^{n}\sum_{t=1}^{T}E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it}+\theta_t)\right)\times \\
&\prod_{j=0}^{p}\exp\left(\frac{-\beta_j^2}{2000}\right).
\end{aligned}$$

$\boldsymbol{\beta}=(\beta_0,\ldots,\beta_p)$ is drawn as a block for all p covariates, including the intercept term β_0, via a Metropolis-Hastings step.

Update ϕ_{it}

Each ϕ_{it} is sampled separately using a Metropolis-Hastings step. The full conditional distribution for ϕ_{it} is

$$\begin{aligned}
f(\phi_{it}|Y_{it}) &\propto \text{Poisson}(Y_{it}|\phi_{it}) \times \text{N}(\phi_{it}|\boldsymbol{\phi}_{-it}) \\
&\propto \left(E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it}+\theta_t)\right)^{Y_{it}}\exp\left(-E_{it}\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it}+\theta_t\right)\right)\times \\
&\text{N}\left(\frac{\rho_s\sum_{j=1}^{n}w_{ij}\phi_{jt}}{\rho_s\sum_{j=1}^{n}w_{ij}+1-\rho_s},\frac{\tau_t^2}{\rho_s\sum_{j=1}^{n}w_{ij}+1-\rho_s}\right)
\end{aligned}$$

Note, this updating step is outlined for variant A of the model and the step for variant B is analogous by simply replacing ρ_s with ρ_{s_t}.

Update θ_t

The joint distribution for $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_T)$ can be written as

$$f(\theta_1,\ldots,\theta_T)=f(\theta_1)f(\theta_2|\theta_1)f(\theta_3|\theta_2,\theta_1)\ldots f(\theta_T|\theta_{T-1},\ldots,\theta_1).$$

Since $\theta_1\sim\text{N}(0,\sigma^2)$, we get

$$\begin{aligned}
f(\theta_1) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\theta_1^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{\theta_1^2}{2\sigma^2}\right).
\end{aligned}$$

Since $\theta_t|\theta_{t-1} \sim \mathrm{N}\left(\alpha\theta_{t-1}, \sigma^2\right)$, we get

$$\begin{aligned} f(\theta_t|\theta_{t-1}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta_t - \alpha\theta_{t-1})^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\theta_t - \alpha\theta_{t-1})^2}{2\sigma^2}\right). \end{aligned}$$

Therefore the joint probability distribution for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_T)$ is given by

$$\begin{aligned} f(\theta_1, \dots, \theta_T) &= f(\theta_1) \prod_{t=2}^{T} f(\theta_t|\theta_{t-1}) \\ &\propto \exp\left(-\frac{\theta_1^2}{2\sigma^2}\right) \prod_{t=2}^{T} \exp\left(-\frac{(\theta_t - \alpha\theta_{t-1})^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\theta_1^2}{\sigma^2} + \sum_{t=2}^{T} \frac{(\theta_t - \alpha\theta_{t-1})^2}{\sigma^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{R} \boldsymbol{\theta}\right) \\ &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{R}^{-1}), \end{aligned}$$

where $\boldsymbol{R}$ is a $T \times T$ precision matrix, with element R_{ts} given as

$$R_{ts} = \begin{cases} \frac{1+\alpha^2}{\sigma^2}, & \text{if } t = s \neq T, \\ \frac{1}{\sigma^2}, & \text{if } t = s = T, \\ -\frac{\alpha}{\sigma^2}, & \text{if } |t - s| = 1, \\ 0, & \text{if } |t - s| \geq 2. \end{cases}$$

According to multivariate Gaussian theory, the conditional expectation of the distribution $\theta_t|\boldsymbol{\theta}_{-t} \sim \mathrm{N}\left(\mathrm{E}[\theta_t|\boldsymbol{\theta}_{-t}], \mathrm{Var}[\theta_t|\boldsymbol{\theta}_{-t}]\right)$ is

$$\mathrm{E}[\theta_t|\boldsymbol{\theta}_{-t}] = \begin{cases} \frac{\alpha}{1+\alpha^2}\theta_{t+1}, & \text{if } t = 1, \\ \frac{\alpha}{1+\alpha^2}\left(\theta_{t-1} + \theta_{t+1}\right), & \text{if } t = 2, \dots, T-1, \\ \alpha\theta_{t-1}, & \text{if } t = T. \end{cases}$$

The conditional variance $\mathrm{Var}[\theta_t|\boldsymbol{\theta}_{-t}]$ is

$$\mathrm{Var}[\theta_t|\boldsymbol{\theta}_{-t}] = \begin{cases} \frac{\sigma^2}{1+\alpha^2}, & \text{if } t \neq T, \\ \sigma^2, & \text{if } t = T. \end{cases}$$

Each θ_t is sampled via a Metropolis-Hastings step. The full conditional distribution for θ_t is

$$f(\theta_t|\boldsymbol{Y}) \propto \prod_{i=1}^{n} \text{Poisson}(Y_{it}|\theta_t) \times \text{N}\left(\text{E}[\theta_t|\boldsymbol{\theta}_{-t}], \text{Var}[\theta_t|\boldsymbol{\theta}_{-t}]\right).$$

When $t = 1$:

$$f(\theta_t|\boldsymbol{Y}) \propto \left(\prod_{i=1}^{n}\left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right)^{Y_{it}}\right) \exp\left(-\sum_{i=1}^{n} E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right) \times \exp\left(\frac{\left(\theta_t - \frac{\alpha}{1+\alpha^2}\theta_{t+1}\right)^2}{-2\frac{\sigma^2}{1+\alpha^2}}\right).$$

When $t = 2, \dots, T-1$:

$$f(\theta_t|\boldsymbol{Y}) \propto \left(\prod_{i=1}^{n}\left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right)^{Y_{it}}\right) \exp\left(-\sum_{i=1}^{n} E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right) \times \exp\left(\frac{\left(\theta_t - \frac{\alpha}{1+\alpha^2}\left(\theta_{t-1} + \theta_{t+1}\right)\right)^2}{-2\frac{\sigma^2}{1+\alpha^2}}\right).$$

When $t = T$:

$$f(\theta_t|\boldsymbol{Y}) \propto \left(\prod_{i=1}^{n}\left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right)^{Y_{it}}\right) \exp\left(-\sum_{i=1}^{n} E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it} + \theta_t)\right) \times \exp\left(\frac{\left(\theta_t - \alpha\theta_{t-1}\right)^2}{-2\sigma^2}\right).$$

Update α

α is drawn using a Gibbs sampler. The full conditional probability distribution for α is

$$\begin{aligned} f(\alpha|\boldsymbol{\theta}) &\propto \prod_{t=2}^{T} \text{N}(\theta_t|\alpha\theta_{t-1}, \sigma^2) \times \text{Uniform}(0,1) \\ &\propto \prod_{t=2}^{T} \exp\left(\frac{-(\theta_t - \alpha\theta_{t-1})^2}{2\sigma^2}\right) \\ &\propto \exp\left(\sum_{t=2}^{T} \frac{-(\theta_t - \alpha\theta_{t-1})^2}{2\sigma^2}\right) \\ &\sim \text{N}(m, v), \end{aligned}$$

where $m = \frac{\sum_{t=2}^{T}\theta_t\theta_{t-1}}{\sum_{t=2}^{T}\theta_{t-1}^2}$, and $v = \frac{\sigma^2}{\sum_{t=2}^{T}\theta_{t-1}^2}$.

Update σ^2

σ^2 is drawn using a Gibbs sampler. The full conditional probability distribution for σ^2 is

$$
\begin{aligned}
f(\sigma^2|\boldsymbol{\theta}) &\propto \mathrm{N}\left(\theta_1|0,\sigma^2\right)\prod_{t=2}^{T}\mathrm{N}\left(\theta_t-\alpha\theta_{t-1}|0,\sigma^2\right)\times \text{Inverse-Gamma}(1,0.01)\\
&\sim \text{Inverse-Gamma}(\tilde{a},\tilde{b}),
\end{aligned}
$$

where $\tilde{a}=1+\frac{T}{2}$, and $\tilde{b}=0.01+\frac{1}{2}\left(\theta_1^2+\sum_{t=2}^{T}(\theta_t-\alpha\theta_{t-1})^2\right)$. σ^2 is evaluated at each iteration of Gibbs sampling by directly drawing samples from the above Inverse-Gamma distribution.

Update τ_t^2

τ_t^2 is drawn using a Gibbs sampler. The full conditional distribution for τ_t^2 is

$$
\begin{aligned}
f(\tau_t^2|\boldsymbol{\phi}_t) &\propto \mathrm{N}\left(\boldsymbol{\phi}_t|\mathbf{0},\tau_t^2\boldsymbol{Q}(\rho_s,\widetilde{\boldsymbol{W}})^{-1}\right)\times \text{Inverse-Gamma}(1,0.01)\\
&\sim \text{Inverse-Gamma}(\tilde{a},\tilde{b}),
\end{aligned}
$$

where $\tilde{a}=1+\frac{n}{2}$ and $\tilde{b}=0.01+\frac{1}{2}\boldsymbol{\phi}_t^\top\boldsymbol{Q}(\rho_s,\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_t$. τ_t^2 is evaluated at each iteration of Gibbs sampling by directly drawing samples from the above Inverse-Gamma distribution. Note, this updating step is outlined for variant A of the model and the step for variant B is analogous by replacing ρ_s with ρ_{s_t} and $\widetilde{\boldsymbol{W}}$ with $\widetilde{\boldsymbol{W}}_t$.

Update $\widetilde{\boldsymbol{W}}$ or $\widetilde{\boldsymbol{W}}_t$

$\widetilde{\boldsymbol{W}}$ or $\widetilde{\boldsymbol{W}}_t$ is sampled via a Metropolis-Hastings step. For variant A of the model, the full conditional distribution for $\widetilde{\boldsymbol{W}}$ is

$$
\begin{aligned}
f(\widetilde{\boldsymbol{W}}|\boldsymbol{\phi}) &\propto \prod_{t=1}^{T}\mathrm{N}\left(\boldsymbol{\phi}_t|\mathbf{0},\tau_t^2\boldsymbol{Q}(\rho_s,\widetilde{\boldsymbol{W}})^{-1}\right)\times f\left(\widetilde{\boldsymbol{W}}=\boldsymbol{W}^{(c,k)}\right)\\
&\propto ||\boldsymbol{Q}(\rho_s,\widetilde{\boldsymbol{W}})||^{\frac{T}{2}}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{\phi}_t^\top\boldsymbol{Q}(\rho_s,\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_t\right)\tau_t^{-2}\right),
\end{aligned}
$$

where $||\cdot||$ denotes the determinant of a matrix. For variant B of the model, the full conditional distribution for $\widetilde{\boldsymbol{W}}_t$ is

$$
\begin{aligned}
f(\widetilde{\boldsymbol{W}}_t|\boldsymbol{\phi}_t) &\propto \mathrm{N}\left(\boldsymbol{\phi}_t|\mathbf{0},\tau_t^2\boldsymbol{Q}(\rho_{s_t},\widetilde{\boldsymbol{W}}_t)^{-1}\right)\times f\left(\widetilde{\boldsymbol{W}}_t=\boldsymbol{W}^{(c,k,t)}\right)\\
&\propto ||\boldsymbol{Q}(\rho_{s_t},\widetilde{\boldsymbol{W}}_t)||^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}_t^\top\boldsymbol{Q}(\rho_{s_t},\widetilde{\boldsymbol{W}}_t)\boldsymbol{\phi}_t\right)\tau_t^{-2}\right).
\end{aligned}
$$

Update ρ_s or ρ_{s_t}

For variant A of the model, the full conditional distribution for ρ_s is

$$\begin{aligned} f(\rho_s|\boldsymbol{\phi}) &\propto \prod_{t=1}^{T} \mathrm{N}\left(\boldsymbol{\phi}_t|\mathbf{0}, \tau_t^2 \boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}})^{-1}\right) \times \mathrm{Uniform}(0,1) \\ &\propto ||\boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}})||^{\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{\phi}_t^{\top}\boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}})\boldsymbol{\phi}_t\right)\tau_t^{-2}\right). \end{aligned}$$

For variant B of the model, the full conditional distribution for ρ_{s_t} is

$$\begin{aligned} f(\rho_{s_t}|\boldsymbol{\phi}_t) &\propto \mathrm{N}\left(\boldsymbol{\phi}_t|\mathbf{0}, \tau_t^2 \boldsymbol{Q}(\rho_s, \widetilde{\boldsymbol{W}}_t)^{-1}\right) \times \mathrm{Uniform}(0,1) \\ &\propto ||\boldsymbol{Q}(\rho_{s_t}, \widetilde{\boldsymbol{W}}_t)||^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}_t^{\top}\boldsymbol{Q}(\rho_{s_t}, \widetilde{\boldsymbol{W}}_t)\boldsymbol{\phi}_t\right)\tau_t^{-2}\right). \end{aligned}$$

5.4 Simulation

In this section a simulation study is presented to comprehensively quantify the performance of the proposed methodology. The study assesses the performance of both model variants (A - static and B - time varying), and in both cases I compare models where the global spatial dependence parameters (ρ_s, ρ_{s_t}) are fixed at 0.99 or estimated within the model. Thus in the study five different models are compared, where **ST-A** and **ST-B** denote model variants A and B where $\rho_s, \rho_{s_t} = 0.99$, while **ST-A*** and **ST-B*** denote model variants A and B where (ρ_s, ρ_{s_t}) are estimated within the model. Finally, model **ST-N** is the existing non-cluster model proposed by Napier et al. (2016) (see Section 5.2.2). The aims in this study are to illustrate the improved risk estimation delivered by our models compared to a similar non-clustering alternative, and also to quantify the accuracy of the resulting cluster identification.

5.4.1 Data generation

Simulated disease count data $\{Y_{it}\}$ are generated from the Poisson log-linear model (5.7) for the $n = 257$ IZs that comprise the Greater Glasgow study region for $T = 7$ time periods. The template for the expected disease counts $\{E_{it}\}$ is based on the motivating study data, whose values range between 12.61 and 160.15 in a single IZ and year with a median of 74.09. However, to explore the impact of disease prevalence on model performance, these $\{E_{it}\}$ values are divided by the scale factors (SF) of 1, 2 and 4. Thus SF = 1 corresponds to the motivating study data, SF = 2 corresponds to having a smaller number of expected counts, while SF = 4 represents a rare disease that has very small expected counts.

Disease risks $\{R_{it}\}$ are generated by simulating both spatial $(\{\phi_{it}\})$ and temporal $(\{\theta_t\})$ random effects, and as previously described covariates are not included. The temporal random effects are generated from the Gaussian AR(1) model given by (5.8), where we fix $\alpha = 0.9$ and $\sigma^2 = 0.1$. The spatially correlated random effects for each time period $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{nt})$ are generated from a multivariate Gaussian distribution with the spatially correlated precision matrix proposed by Leroux et al. (2000) where τ_t^2 is fixed at 0.001 for each t. To assess model performance with different degrees of spatial correlation in the risk surface, the spatially correlated random effects $\boldsymbol{\phi}_t$ are generated by varying the spatial dependence parameters over $\rho_s, \rho_{s_t} = 0.9, 0.6, 0.3, 0$. Here a value of 0.9 corresponds to strong spatial dependence, values of (0.6, 0.3) correspond to moderate and weak dependence respectively while a value of 0 corresponds to spatial independence.

Clustering is induced into these spatial surfaces by the mean of the multivariate Gaussian distribution used to generate $\boldsymbol{\phi}_t$, which is denoted by $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{nt})$. At each time period we consider high, medium and low risk levels, which are generated by specifying a piecewise constant mean function so that each $\mu_{it} \in \{-Z, 0, Z\}$. Thus geographically neighbouring areal units that have the same mean value are in the same cluster as their disease risks will be similar, while those pairs that have different mean values will exhibit a step-change in their risks and hence are in different clusters. The value of Z is varied in the different scenarios of our simulation design as either $Z = 0.5$ or $Z = 1$, which respectively correspond to small and large differences in disease risk between neighbouring areal units in different clusters. Two cases are considered for this clustering, which respectively favour variants A (Case 1) and B (Case 2) of our model.

- **Case 1** - the simulated clusters remain constant during the study period, which is achieved by setting $\mu_{it} = \mu_{il}$ for all $t \neq l$.
- **Case 2** - the simulated clusters evolve over time, which is achieved by $\mu_{it} \neq \mu_{il}$ for some pairs of time periods (t, l).

Figure 5.2 provides maps of the cluster structures simulated for both Case 1 and Case 2, where areal units in the high-risk, medium-risk and low-risk clusters are respectively shaded in black, grey and white. Under Case 1 the simulated clusters are common to all time periods, while under Case 2 the cluster structures at time periods $t = 1, 4, 7$ are shown here. For Case 2 the temporal evolution of the clusters is achieved by randomly selecting a small number of areal units and changing their cluster membership for each time period. The chosen cluster structure template is based on the motivating study, by partitioning the set of IZs into three groups based on their SIR values. Finally, a scenario where $Z = 0$ is also considered, which corresponds to a spatially smooth risk surface with no clusters for each time period. Note, in this case as there are no clusters in the simulated risk surfaces then there is no difference between Case 1 and Case 2. The simulation study thus has 30 sub-scenarios in its design, which are summarised in Table 5.1. Thus in this study we vary the following quantities: (i) constant and time-varying clusters via Case 1 and Case 2; (ii) varying cluster magnitudes via $Z = 1, 0.5, 0$; (iii) varying disease prevalences via $\text{SF} = 1, 2, 4$; and (iv) varying levels of spatial autocorrelation via $\rho_s, \rho_{s_t} = 0.9, 0.6, 0.3, 0$.

Table 5.1: Description of the scenarios in the simulation study. SF indicates the scale factor used for the expected values.

Cluster cases	Z	SF	(ρ_s, ρ_{s_t})
Case 1 / 2	$Z \in \{1, 0.5\}$	$\text{SF} \in \{1, 2, 4\}$	$\rho_s, \rho_{s_t} = 0.9$
- -	$Z = 0$	$\text{SF} \in \{1, 2, 4\}$	$\rho_s, \rho_{s_t} = 0.9$
Case 1 / 2	$Z \in \{1, 0.5\}$	$\text{SF} = 1$	$\rho_s, \rho_{s_t} \in \{0.6, 0.3, 0\}$
- -	$Z = 0$	$\text{SF} = 1$	$\rho_s, \rho_{s_t} \in \{0.6, 0.3, 0\}$

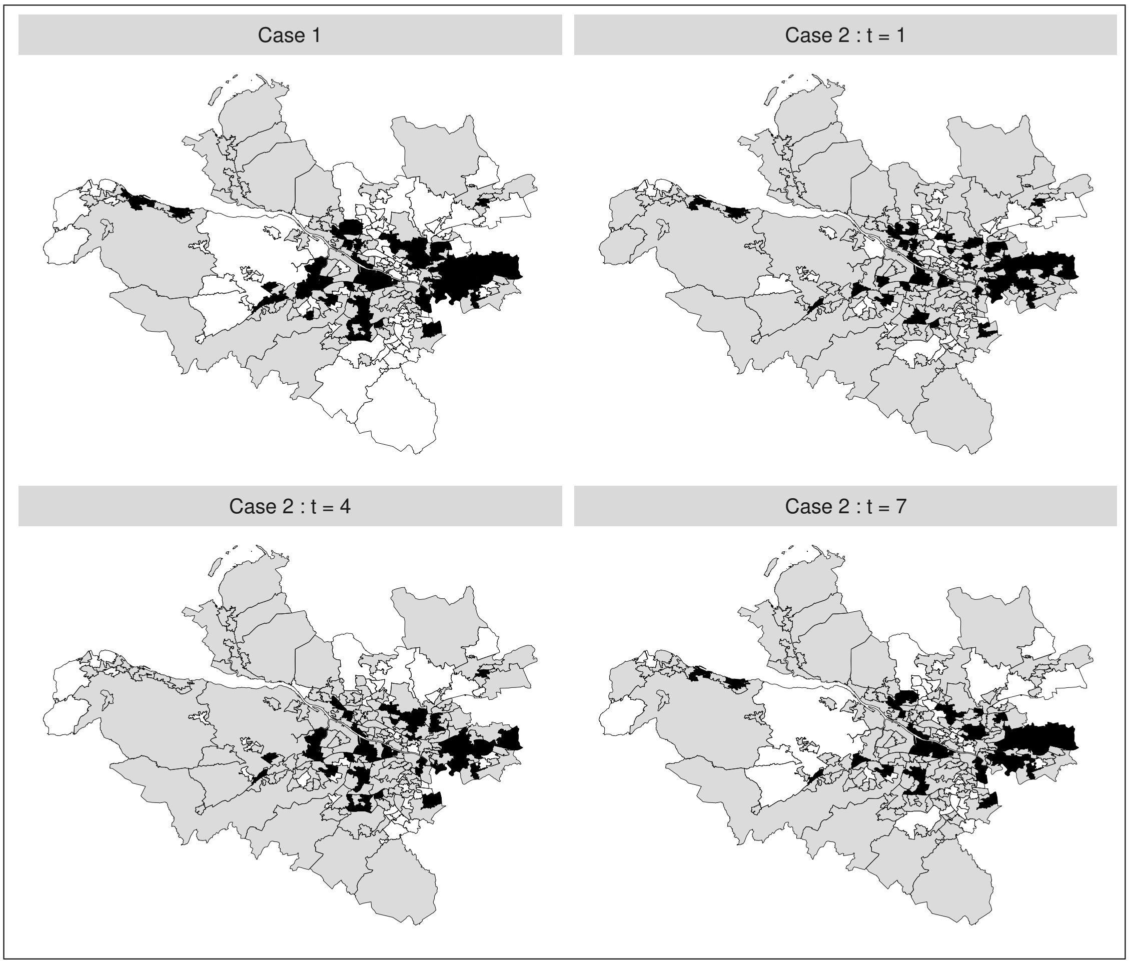


Figure 5.2: Maps of the simulated cluster structures under Case 1 (top-left) and Case 2 (top-right and bottom). High-risk, medium-risk and low-risk clusters are respectively shaded in black, grey and white.

5.4.2 Results

One hundred simulated data sets are generated under each of the 30 scenarios shown in Table 5.1, and the five models **ST-A**, **ST-A***, **ST-B**, **ST-B*** and **ST-N** are fitted to each data set. In all cases inference is based on a single Markov chain with 100,000 MCMC samples, 80,000 of which were discarded for the burn-in period and the remaining 20,000 samples were thinned by 10 to reduce their autocorrelation. The main results are shown below, while the sensitivity of these results to the choice of prior distribution is assessed in Section 5.5.

The results are shown in Tables 5.2 and 5.3, which summarise the modelling performance using three metrics. The accuracy of disease risk estimation is quantified by the root mean square error (RMSE) of the risk estimates and the corresponding coverage probabilities of the 95% credible intervals. The correctness of the estimated cluster structures is measured by the adjusted Rand Index (Hubert and Arabie, 1985) between the true and estimated cluster structures over all time periods. For this metric a value of 1 indicates complete agreement between two cluster structures, a value of 0 indicates that the data points are randomly allocated to the two cluster structures, and a value less than 0 indicates that the level of agreement between the two cluster structures is smaller than that arising from randomly allocated data points. Table 5.2 displays these performance metrics for each model under different scenarios of Cases 1 / 2, $Z = 1, 0.5, 0$ and SF $= 1, 2, 4$, but in this table the spatial random effects are simulated under the strong spatial dependence scenario with $\rho_s, \rho_{s_t} = 0.9$.

The non-cluster model **ST-N** mainly performs poorly in terms of risk estimation compared to the clustering models proposed here, having mostly larger RMSE values and coverage probabilities as low as 0.8. An exception to this is when there are no clusters in disease risk ($Z = 0$), and in this case the **ST-N** model performs similarly to the constant cluster models **ST-A** and **ST-A***. The other scenario in which **ST-N** performs comparably to the cluster models is when the disease is rare (SF $= 4$) and the cluster boundaries are small in magnitude ($Z = 0.5$), which is because in this scenario the clusters are hard to identify based on their small size and small numbers of disease cases.

The four clustering models perform best when the disease prevalence is high (SF $=$ 1) and the clusters are large in size ($Z = 1$) as expected, which is because in these scenarios there are more disease cases from which to identify more prominent clusters. In this situation the RMSE values are very low compared to the range of the true risks, coverage probabilities are close to their nominal 0.95 levels, and cluster identification is very good with ARI values either very close to or equal to one. When the disease prevalence decreases (SF > 1) and the clusters are less pronounced the models perform less well, but the ARI values are still relatively close to one in most scenarios unless the disease is rare (SF $=$ 4) and the clusters are small in magnitude ($Z = 0.5$). Additionally, the models identify temporally static clusters better than temporally varying ones, as the results for models **ST-A** and **ST-A*** in Case 1 are generally better than those for models **ST-B** and **ST-B*** in Case 2. This improved performance in the static cluster case is because the temporal replication in the true cluster structure yields more data from which to identify clusters, thus resulting in improved model performance.

When the clusters are temporally constant (Case 1) then as expected models **ST-A** and **ST-A***, which assume static clusters, produce more accurate risk estimates (lower RMSE) than models **ST-B** and **ST-B*** in all cases, as well as producing adjusted Rand indices that are generally very close to one. Similarly, when the clusters evolve over time (Case 2) then as expected models **ST-B** and **ST-B***, which allow for dynamic clusters, perform better than **ST-A** and **ST-A*** in most scenarios. The only slight exception to this is when the disease is rare (SF $= 4$) and the clusters are not large in size ($Z = 0.5$), which is the case where the clusters are hardest to identify and hence all the cluster models perform less well (small ARI values). Estimating (ρ_s, ρ_{s_t}) (**ST-A***, **ST-B***) rather than fixing them at 0.99 (**ST-A**, **ST-B**) produces better results overall, with lower RMSE values and higher ARI values in almost all scenarios.

Table 5.3 summarises model performance under different levels of spatial autocorrelation (values of ρ_s, ρ_{s_t}), and the results are based on the expected disease counts from the motivating data (SF $= 1$). The results show that reducing the spatial dependence in the data does not seem to have any substantial effect on the ability of the proposed cluster models to estimate disease risk or identify the correct cluster structure, as the ARI results do not differ greatly as the spatial dependence parameters (ρ_s, ρ_{s_t}) vary. Additionally, the coverage probabilities are also largely unaffected by this change, and the RMSE values increase very slightly (suggesting worse estimation) as (ρ_s, ρ_{s_t}) gets closer to zero, due to a reduction in the borrowing of strength spatially when doing the estimation. In Case 1 **ST-A*** again outperforms **ST-A** across the board, suggesting that estimating ρ_s leads to better model performance regardless of the level of spatial autocorrelation in the data. This effect is also seen when comparing **ST-B*** and **ST-B** in Case 2 in terms of RMSE, but for the ARI results the two models are very similar. Finally, the globally smooth non-clustering model **ST-N** again performs uniformly badly when there are clusters in the data ($Z > 0$) as expected.

Table 5.2: Median values of the RMSE, 95% credible interval coverages and adjusted Rand Index (ARI) for each model and scenario. In the scenarios considered here $\rho_s = \rho_{s_t} = 0.9$ is used to simulate the spatial random effects for each time period.

Performance metric	Cluster case	Z	SF	Model				
				ST-A	**ST-A***	**ST-B**	**ST-B***	**ST-N**
RMSE	**Case 1**	1	1	0.087	0.068	0.091	0.074	1.175
		0.5	1	0.071	0.060	0.099	0.096	0.119
		1	2	0.115	0.094	0.130	0.116	1.155
		0.5	2	0.097	0.082	0.159	0.158	0.157
		1	4	0.150	0.128	0.207	0.192	1.109
		0.5	4	0.151	0.120	0.225	0.224	0.204
	Case 2	1	1	0.135	0.132	0.091	0.074	0.904
		0.5	1	0.114	0.111	0.101	0.099	0.113
		1	2	0.188	0.181	0.142	0.126	0.864
		0.5	2	0.153	0.148	0.155	0.155	0.150
		1	4	0.267	0.249	0.217	0.209	0.817
		0.5	4	0.204	0.196	0.218	0.217	0.195
	- -	0	1	0.025	0.034	0.083	0.070	0.026
		0	2	0.029	0.040	0.112	0.095	0.028
		0	4	0.091	0.049	0.145	0.131	0.033
Coverage probability	**Case 1**	1	1	0.975	0.975	0.969	0.955	0.801
		0.5	1	0.974	0.974	0.928	0.920	0.951
		1	2	0.979	0.973	0.948	0.931	0.805
		0.5	2	0.972	0.974	0.843	0.842	0.949
		1	4	0.981	0.969	0.907	0.897	0.843
		0.5	4	0.937	0.959	0.777	0.764	0.947
	Case 2	1	1	0.935	0.942	0.964	0.969	0.811
		0.5	1	0.937	0.930	0.900	0.884	0.950
		1	2	0.915	0.938	0.921	0.923	0.851
		0.5	2	0.928	0.913	0.827	0.819	0.948
		1	4	0.862	0.929	0.885	0.860	0.881
		0.5	4	0.901	0.878	0.783	0.785	0.942
	- -	0	1	0.991	0.993	0.694	0.905	0.996
		0	2	0.992	0.997	0.759	0.893	0.998
		0	4	0.979	0.998	0.829	0.891	0.999
Adjusted Rand Index (ARI)	**Case 1**	1	1	1	1	0.986	0.976	- -
		0.5	1	1	1	0.843	0.853	- -
		1	2	1	1	0.891	0.893	- -
		0.5	2	0.983	1	0.514	0.589	- -
		1	4	1	1	0.740	0.787	- -
		0.5	4	0.617	0.937	0.319	0.367	- -
	Case 2	1	1	0.353	0.409	0.987	0.987	- -
		0.5	1	0	0.412	0.711	0.655	- -
		1	2	0.351	0.393	0.608	0.907	- -
		0.5	2	0	0.360	0.383	0.426	- -
		1	4	0.359	0.395	0.504	0.600	- -
		0.5	4	0	0.290	0.235	0.267	- -
	- -	0	1	1	0	0	0	- -
		0	2	1	0	0	0	- -
		0	4	0	0	0	0	- -

Table 5.3: Median values of the RMSE, 95% credible interval coverages and adjusted Rand Index (ARI) for each model and scenario. In the scenarios considered here the expected counts from the motivating application are used to generate disease data.

Performance metric	Cluster case	Z	ρ_s/ρ_{s_t}	Model				
				ST-A	**ST-A***	**ST-B**	**ST-B***	**ST-N**
RMSE	**Case 1**	1	0.9	0.087	0.068	0.091	0.074	1.175
		0.5	0.9	0.071	0.060	0.099	0.096	0.119
		1	0.6	0.088	0.068	0.090	0.074	1.172
		0.5	0.6	0.076	0.059	0.097	0.094	0.117
		1	0.3	0.089	0.071	0.092	0.077	1.191
		0.5	0.3	0.100	0.061	0.099	0.095	0.118
		1	0	0.094	0.078	0.097	0.084	1.246
		0.5	0	0.102	0.064	0.101	0.100	0.117
	Case 2	1	0.9	0.135	0.132	0.091	0.074	0.904
		0.5	0.9	0.114	0.111	0.101	0.099	0.113
		1	0.6	0.134	0.128	0.089	0.075	0.892
		0.5	0.6	0.114	0.110	0.102	0.100	0.113
		1	0.3	0.131	0.128	0.092	0.076	0.868
		0.5	0.3	0.116	0.113	0.101	0.101	0.115
		1	0	0.134	0.129	0.094	0.081	0.884
		0.5	0	0.116	0.112	0.105	0.103	0.115
	- -	0	0.9	0.025	0.034	0.083	0.070	0.026
		0	0.6	0.026	0.030	0.084	0.071	0.024
		0	0.3	0.029	0.032	0.085	0.071	0.026
		0	0	0.035	0.038	0.085	0.073	0.033
Coverage probability	**Case 1**	1	0.9	0.979	0.975	0.969	0.955	0.801
		0.5	0.9	0.974	0.974	0.928	0.920	0.951
		1	0.6	0.978	0.971	0.969	0.954	0.791
		0.5	0.6	0.973	0.972	0.930	0.925	0.952
		1	0.3	0.978	0.968	0.968	0.952	0.797
		0.5	0.3	0.947	0.969	0.922	0.921	0.949
		1	0	0.974	0.946	0.964	0.934	0.781
		0.5	0	0.947	0.950	0.909	0.899	0.949
	Case 2	1	0.9	0.935	0.942	0.964	0.969	0.811
		0.5	0.9	0.937	0.930	0.900	0.884	0.950
		1	0.6	0.932	0.943	0.961	0.967	0.809
		0.5	0.6	0.935	0.929	0.898	0.889	0.949
		1	0.3	0.934	0.943	0.957	0.962	0.810
		0.5	0.3	0.936	0.930	0.900	0.894	0.949
		1	0	0.936	0.943	0.951	0.941	0.793
		0.5	0	0.935	0.930	0.877	0.869	0.949
	- -	0	0.9	0.991	0.993	0.694	0.905	0.996
		0	0.6	0.987	0.997	0.710	0.884	0.998
		0	0.3	0.977	0.995	0.683	0.885	0.996
		0	0	0.934	0.984	0.705	0.861	0.982
Adjusted Rand Index (ARI)	**Case 1**	1	0.9	1	1	0.986	0.976	- -
		0.5	0.9	1	1	0.843	0.853	- -
		1	0.6	1	1	0.985	0.975	- -
		0.5	0.6	0.992	1	0.847	0.846	- -
		1	0.3	1	1	0.985	0.975	- -
		0.5	0.3	0.541	1	0.846	0.856	- -
		1	0	1	1	0.988	0.976	- -
		0.5	0	0.541	1	0.828	0.839	- -
	Case 2	1	0.9	0.353	0.409	0.987	0.987	- -
		0.5	0.9	0	0.412	0.711	0.655	- -
		1	0.6	0.367	0.412	0.987	0.988	- -
		0.5	0.6	0	0.411	0.645	0.699	- -
		1	0.3	0.358	0.409	0.980	0.987	- -
		0.5	0.3	0	0.388	0.691	0.735	- -
		1	0	0.357	0.397	0.981	0.987	- -
		0.5	0	0	0.387	0.618	0.687	- -
	- -	0	0.9	1	0	0	0	- -
		0	0.6	1	0	0	0	- -
		0	0.3	1	0	0	0	- -
		0	0	1	0	0	0	- -

5.5 Sensitivity analysis to changing the prior distribution for τ_t^2

In Section 5.4 an Inverse-Gamma$(1,0.01)$ prior is specified for the spatial random effects variance τ_t^2 in the proposed models. To assess the impact of this prior for τ_t^2 on model performance, I re-run part of the simulation study by fitting the clustering models separately with both Inverse-Gamma$(0.001,0.001)$ and Inverse-Gamma$(0.5,0.0005)$ priors, which are two commonly used alternatives in the literature (see Anderson et al. (2014) and Spiegelhalter et al. (1996)). One hundred simulated data sets are generated as described in Section 5.4.1, where $Z = 1,0.5,0$ and both Cases 1 (static clusters) and 2 (time-varying clusters) are considered. In generating the data $(\rho_s,\ \rho_{s_t})$ are fixed at 0.9, and the expected number of disease cases from the motivating data are used (i.e. SF = 1). The proposed models **ST-A**, **ST-B**, **ST-A*** and **ST-B*** are respectively applied to the data using the three different choices of prior Inverse-Gamma distribution for τ_t^2, and the results are summarised in Table 5.4.

The results show that changing the hyperparameters of the Inverse-Gamma prior for τ_t^2 does not seem to have any substantial effect on the ability of the cluster models to estimate disease risk or identify the correct cluster structure, as the differences in RMSE, 95% coverage probabilities and ARI values are very minimal when the prior varies. When the clusters are temporally constant (Case 1) **ST-A** and **ST-A*** generally produce lower RMSE values and higher ARI values (very close to one) than models **ST-B** and **ST-B***, excepting the scenario when $Z = 0.5$ and Inverse-Gamma$(0.001,0.001)$ is used. When the clusters evolve over time (Case 2) **ST-B** and **ST-B*** perform better than **ST-A** and **ST-A***. When there are no clusters in disease risk ($Z = 0$), models **ST-A** and **ST-A*** produce lower RMSE values than **ST-B** and **ST-B*** regardless of the choice of the prior, and **ST-A** is the best of the four in terms of cluster identification, with a median ARI of 1. In addition, estimating (ρ_s,ρ_{s_t}) (**ST-A***, **ST-B***) rather than fixing them at 0.99 (**ST-A**, **ST-B**) produces better results overall in terms of both risk estimation and cluster identification in almost all scenarios. These conclusions are consistent with those provided in the simulation study in Section 5.4. Therefore, our methodology appears to be robust to the choice of the hyperparameters of the prior Inverse-Gamma distribution for τ_t^2.

Table 5.4: Median values of the RMSE, 95% credible interval coverages of the risk estimates and adjusted Rand Index (ARI) for each model and scenario.

Performance metric	Cluster case	Z	Inverse-Gamma (IG) prior	Model			
				ST-A	**ST-A***	**ST-B**	**ST-B***
RMSE	**Case 1**	1	IG(1,0.01)	0.088	0.070	0.090	0.075
		1	IG(0.001,0.001)	0.088	0.067	0.090	0.074
		1	IG(0.5,0.0005)	0.088	0.066	0.090	0.072
		0.5	IG(1,0.01)	0.074	0.059	0.098	0.095
		0.5	IG(0.001,0.001)	0.100	0.057	0.099	0.095
		0.5	IG(0.5,0.0005)	0.073	0.056	0.099	0.095
	Case 2	1	IG(1,0.01)	0.132	0.128	0.090	0.076
		1	IG(0.001,0.001)	0.132	0.126	0.091	0.075
		1	IG(0.5,0.0005)	0.132	0.127	0.090	0.076
		0.5	IG(1,0.01)	0.115	0.111	0.102	0.101
		0.5	IG(0.001,0.001)	0.115	0.111	0.099	0.102
		0.5	IG(0.5,0.0005)	0.115	0.110	0.102	0.101
	- -	0	IG(1,0.01)	0.024	0.034	0.082	0.071
		0	IG(0.001,0.001)	0.023	0.034	0.070	0.071
		0	IG(0.5,0.0005)	0.023	0.033	0.069	0.072
Coverage probability	**Case 1**	1	IG(1,0.01)	0.975	0.973	0.968	0.954
		1	IG(0.001,0.001)	0.976	0.958	0.970	0.936
		1	IG(0.5,0.0005)	0.976	0.942	0.969	0.919
		0.5	IG(1,0.01)	0.968	0.974	0.927	0.922
		0.5	IG(0.001,0.001)	0.939	0.958	0.927	0.893
		0.5	IG(0.5,0.0005)	0.969	0.946	0.926	0.876
	Case 2	1	IG(1,0.01)	0.932	0.942	0.964	0.969
		1	IG(0.001,0.001)	0.933	0.942	0.963	0.952
		1	IG(0.5,0.0005)	0.934	0.942	0.964	0.934
		0.5	IG(1,0.01)	0.930	0.928	0.901	0.887
		0.5	IG(0.001,0.001)	0.930	0.930	0.904	0.845
		0.5	IG(0.5,0.0005)	0.931	0.931	0.899	0.817
	- -	0	IG(1,0.01)	0.989	0.994	0.703	0.911
		0	IG(0.001,0.001)	0.961	0.978	0.720	0.843
		0	IG(0.5,0.0005)	0.925	0.955	0.698	0.803
Adjusted Rand Index (ARI)	**Case 1**	1	IG(1,0.01)	1	1	0.986	0.976
		1	IG(0.001,0.001)	1	1	0.986	0.976
		1	IG(0.5,0.0005)	1	1	0.985	0.975
		0.5	IG(1,0.01)	0.995	1	0.851	0.846
		0.5	IG(0.001,0.001)	0.541	1	0.855	0.841
		0.5	IG(0.5,0.0005)	0.994	1	0.847	0.846
	Case 2	1	IG(1,0.01)	0.367	0.386	0.987	0.987
		1	IG(0.001,0.001)	0.347	0.384	0.987	0.987
		1	IG(0.5,0.0005)	0.367	0.384	0.987	0.987
		0.5	IG(1,0.01)	0	0.390	0.671	0.707
		0.5	IG(0.001,0.001)	0	0.388	0.733	0.687
		0.5	IG(0.5,0.0005)	0	0.389	0.628	0.667
	- -	0	IG(1,0.01)	1	0	0	0
		0	IG(0.001,0.001)	1	0	0	0
		0	IG(0.5,0.0005)	1	0	0	0

5.6 Results from the Greater Glasgow respiratory disease study

5.6.1 Model fitting and inference

The four spatio-temporal models **ST-A**, **ST-A***, **ST-B** and **ST-B*** proposed in Section 5.3 are applied to the Greater Glasgow respiratory disease study introduced in Section 5.2.1, with the aims of estimating the spatio-temporal patterns in disease risk and identifying the locations of any high and low risk clusters. Model **ST-N** without clustering is also fitted to the data, to observe how its model fit compares with the clustering models proposed here. In all cases covariate information is not included in the models, because our aim is to identify clusters in the risk surface rather than clusters in the residual risk surface after covariate adjustment. Posterior inference for all models is based on ten independent Markov chains, where each chain is run for 100,000 samples. The first 80,000 samples from each chain were removed as the burn-in period and the remaining samples were thinned by 10, which yields a total of 20,000 samples across all 10 chains. These MCMC samples were deemed to have converged, which was assessed both by examining parameter trace plots and by Geweke (Geweke, 1992) diagnostics.

5.6.2 Overall model fit

The overall fit of each model is summarised in Table 5.5 by the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and the effective number of independent parameters (p_d). The table shows that the four clustering models proposed here fit the data better than the non-clustering model **ST-N**, as the latter has a DIC value that is higher by between 0.7% and 4.9% than those from the clustering models. The time-varying clustering models **ST-B** and **ST-B*** have lower DIC values than the static clustering models **ST-A** and **ST-A***, while allowing (ρ_s, ρ_{s_t}) to be estimated (**ST-A***, **ST-B***) rather than fixed at 0.99 (**ST-A**, **ST-B**) also produces a better model fit. Additionally, our models are also preferred because they model the data using fewer effective parameters, with reductions in p_d varying between 16.0% and 33.6% compared to **ST-N**. This reduced effective number of parameters is due to a reduction in the spatial random effects variance τ_t^2 for the cluster models, which can be seen clearly in Table 5.6. This reduction in the random effects variation is because by our approach the spatial random effects are only forced to smooth towards their neighbours in the same cluster, i.e. those neighbours that have similar random effects values. In contrast, in model **ST-N** this smoothing is with all neighbouring areas, even those that have very different random effect values which hence inflates the variance.

Table 5.5: Deviance Information Criterion (DIC) and the effective number of independent parameters (p_d) for each model.

	ST-A	**ST-A***	**ST-B**	**ST-B***	**ST-N**
DIC	14 631	14 587	14 272	14 040	14 730
p_d	1 344	1 362	1 268	1 183	1 580

5.6.3 Temporal trends in disease risk

The estimated temporal trends in disease risk from model **ST-B*** is presented in Figure 5.3, because it is the best fitting model in terms of DIC. However, the results from the other models are similar, as the mean absolute differences in the posterior median risk estimates between each pair of models range between 0.009 and 0.083 over all years and IZs. Figure 5.3 displays boxplots of the risk estimates from all the areal units over time, and shows a generally increasing trend in risk. These risk estimates are relative to the expected numbers of hospitalisations computed using national respiratory hospitalisation rates for the whole of Scotland between 2011 and 2017, because national rather than Greater Glasgow rates are used by Public Health Scotland when quantifying disease risk. In 2011 the average risk across Greater Glasgow is 1.10, suggesting that on average respiratory disease risk in Greater Glasgow is about 10% higher than the Scottish average. This rises to 1.28 in the final year 2017, which is thus 28% higher than the overall Scotland average. Thus an elevated risk is observed in Glasgow for the entire time period of this study, which corroborates the well-known *Glasgow effect* (Walsh et al., 2010) which is the phenomenon that Glasgow exhibits some of the poorest health in western Europe.

Table 5.6: Summary of the estimated number of clusters with 95% credible intervals and spatial random effects variance τ_t^2 at each time period.

	Time period						
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$
# clusters							
ST-A	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)
ST-A*	2 (2, 5)	2 (2, 5)	2 (2, 5)	2 (2, 5)	2 (2, 5)	2 (2, 5)	2 (2, 5)
ST-B	4 (2, 7)	5 (3, 9)	4 (4, 9)	4 (3, 8)	4 (2, 7)	3 (3, 8)	4 (2, 7)
ST-B*	5 (4, 6)	6 (3, 8)	6 (4, 8)	5 (4, 8)	5 (3, 5)	5 (3, 7)	5 (4, 9)
ST-N	–	–	–	–	–	–	–
τ_t^2							
ST-A	0.056	0.088	0.069	0.083	0.084	0.083	0.075
ST-A*	0.061	0.081	0.069	0.077	0.078	0.081	0.075
ST-B	0.006	0.004	0.004	0.004	0.003	0.004	0.004
ST-B*	0.004	0.003	0.003	0.003	0.003	0.003	0.003
ST-N	0.265	0.291	0.292	0.268	0.254	0.283	0.297

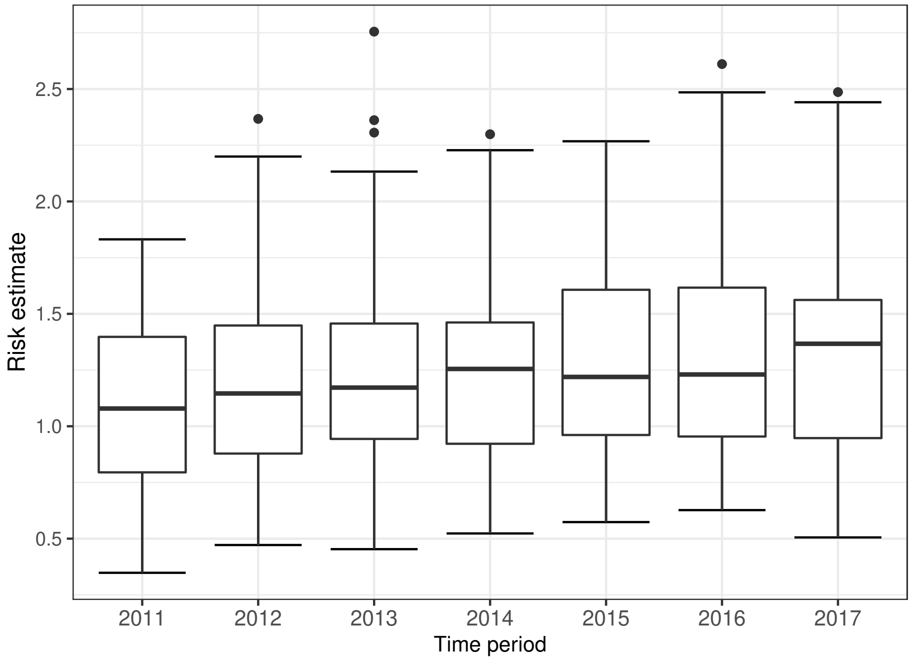


Figure 5.3: Boxplots of the risk estimates (posterior median) from model **ST-B*** for all the areal units over time.

5.6.4 Spatio-temporal cluster structure

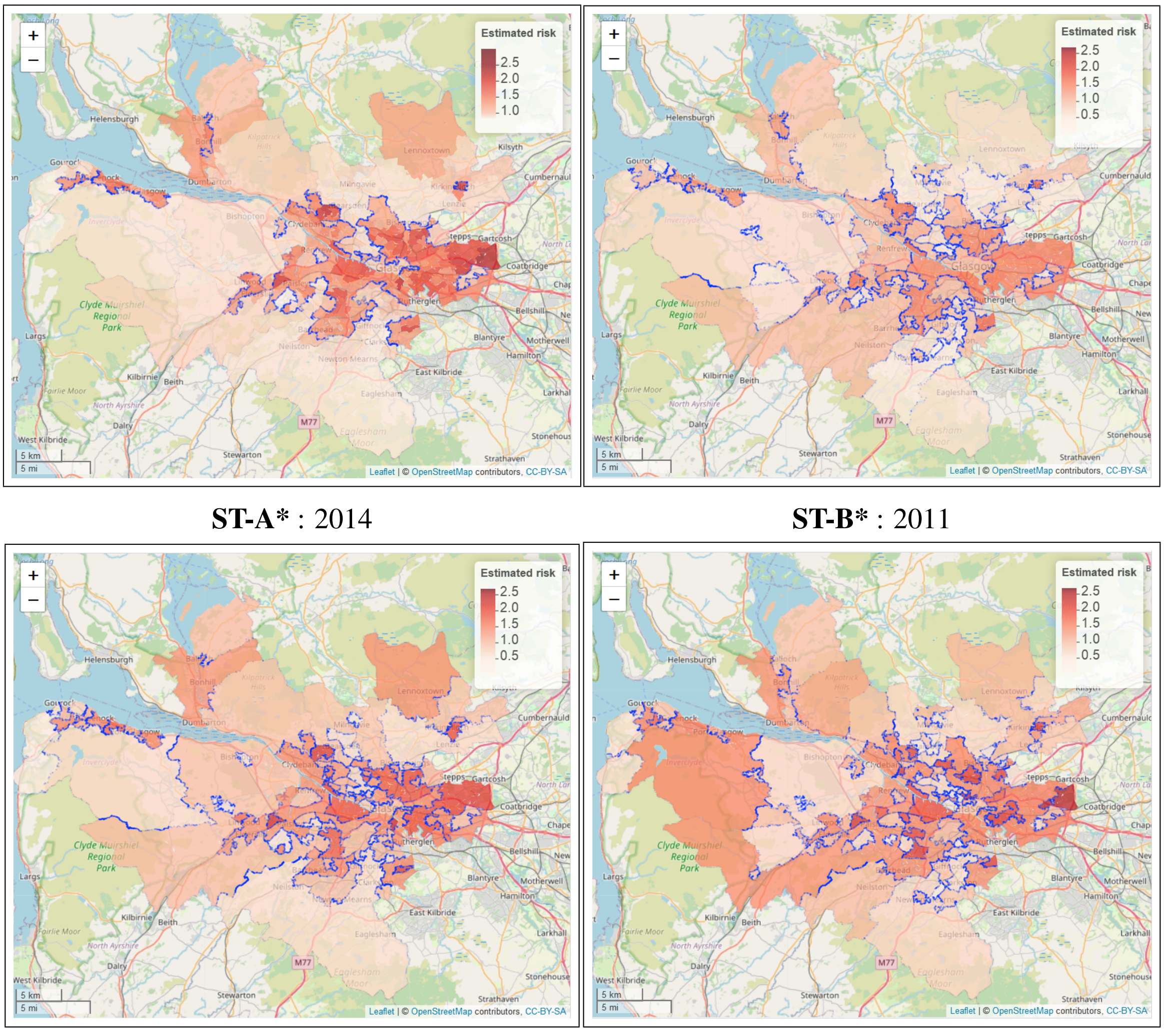
The previous section highlighted that the risk estimates from the clustering and non-clustering models are similar, and so the main advantage of using the clustering models is the additional inference they provide on the locations of clusters of areas that exhibit elevated or reduced disease risks compared to their neighbours. The top panel of Table 5.6 displays the posterior mode for the number of clusters estimated by each cluster model for each year, with uncertainty measured by the 95% credible intervals. If there are multiple modes present in the Markov chain, then the one yielding the fewer effective number of parameters is chosen in the interests of model parsimony. Note that by clusters we mean the number of non-spatial clusters (distinct risk levels) that correspond to the posterior mode of $(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}}_t)$, rather than the number of spatially distinct clusters observable in the risk maps presented below. Models **ST-A** and **ST-A*** have selected the same cluster structure with 2 distinct clusters or risk levels, and that by design this structure is common to all time periods. In contrast, model **ST-B** identifies between 3 and 5 different cluster levels depending on the year, although most years have 4 different clusters, and model **ST-B*** detects 5 distinct clusters for most years. I now present the estimated cluster structures from the models, focusing on both the posterior mode clusters that are static (**ST-A***) and dynamic (**ST-B***) over time, as well as illustrating posterior uncertainty in the cluster structures for **ST-A***.

5.6.4.1 Static and time-varying clusters based on the posterior mode cluster structure

Here I present the results from models **ST-A*** and **ST-B*** as Table 5.5 shows that they have lower DIC values than models **ST-A** and **ST-B**. The top left panel in Figure 5.4 displays the estimated spatial pattern in disease risk from model **ST-A*** for 2014, which is chosen because it is the middle year of the study. The blue dots relate to the cluster boundaries as defined by the posterior mode of $\widetilde{\boldsymbol{W}}$, which in this static clustering case are common to all time periods. The remaining three panels of the figure present the estimated risk surfaces from model **ST-B*** for 2011, 2014 and 2017, the first, middle and last years of the study period. In these cases the clusters/discontinuities denoted by blue dots vary over time and are specific to the year in question, and are again determined by the posterior mode of $\widetilde{\boldsymbol{W}}_t$. These cluster boundaries (discontinuities) represent two IZs that are geographically adjacent but are in different clusters, suggesting they have substantially different risks.

The figure shows that there are a number of similarities between the selected cluster structures from the two cluster models, with the same areas being identified as having very different risks compared to their neighbours. For example, both models identify the large high-risk cluster in the East End of Glasgow (far east of the map), which contains a number of socio-economically deprived areas such as Easterhouse and Barlanark. Additionally, the models identify a cluster of areas to the north of the city containing Springburn and Summerston, as well as another along the southern bank of the River Clyde including Govan and Hutchesontown. They also pick out some high risk areas in the north west including the deprived areas of Drumchapel and Drumry, which are bordered to the north by the more affluent and low-risk Bearsden area. Additionally, another large region of low risk areas identified by both models is the affluent West End of the city such as Dowanhill, which is just to the south of Bearsden. Note, as all the clustering methods were used to adjust the border sharing neighbourhood matrix, clusters cannot be found between areas on opposite banks of the river Clyde, which runs north-west through the study region.

In addition, there is some evidence of a changing cluster structure over time estimated by model **ST-B*** that is worth noting. For example, the large rural areas of Inverclyde in the far west of the study region exhibit low risks in 2011, whereas by 2017 they have joined a moderately high risk cluster. However, comparing the clustering results from the two modelling approaches we find that while **ST-B*** is very flexible in capturing the temporal evolution of clusters, it can also be susceptible to identifying discontinuities (clusters) caused by random noise that are present for some years but not for others, which thus make the interpretation of an evolving cluster structure less clear cut. This happens because in the **ST-B*** model each candidate cluster structure in stage one is elicited using data for a single year, which thus could be affected by random noise in the data. However, model **ST-A*** is less vulnerable to this phenomenon, because the clustering is applied to the data from multiple time periods. Thus despite model **ST-A*** having a higher DIC than model **ST-B***, its consistency of clustering may lead to robust and reliable clusters, as can be visually observed in Figure 5.4. Finally, the cluster discontinuities identified here are determined by the posterior modes of $\widetilde{\boldsymbol{W}}$ from model **ST-A*** and $\widetilde{\boldsymbol{W}}_t$ from **ST-B***. Alternatively, we can make inferential statements about the probability that there is a discontinuity between a certain pair of geographically adjacent areas (i, j), by calculating the probability of element w_{ij} in $\widetilde{\boldsymbol{W}}$ or $\widetilde{\boldsymbol{W}}_t$ being 0 across all the posterior samples.



ST-A* : 2014

ST-B* : 2011

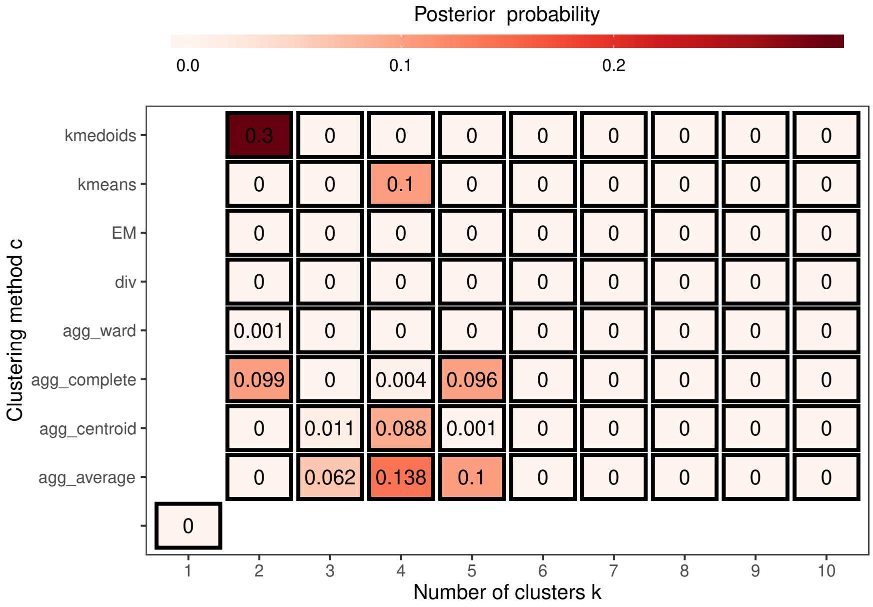
ST-B* : 2014

ST-B* : 2017

Figure 5.4: Maps of the disease risk estimates (posterior median) in Greater Glasgow for 2011, 2014 and 2017 from models **ST-A*** and **ST-B***. The dots on the map indicate the identified cluster discontinuities, which are determined using the posterior mode of $(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}}_t)$.

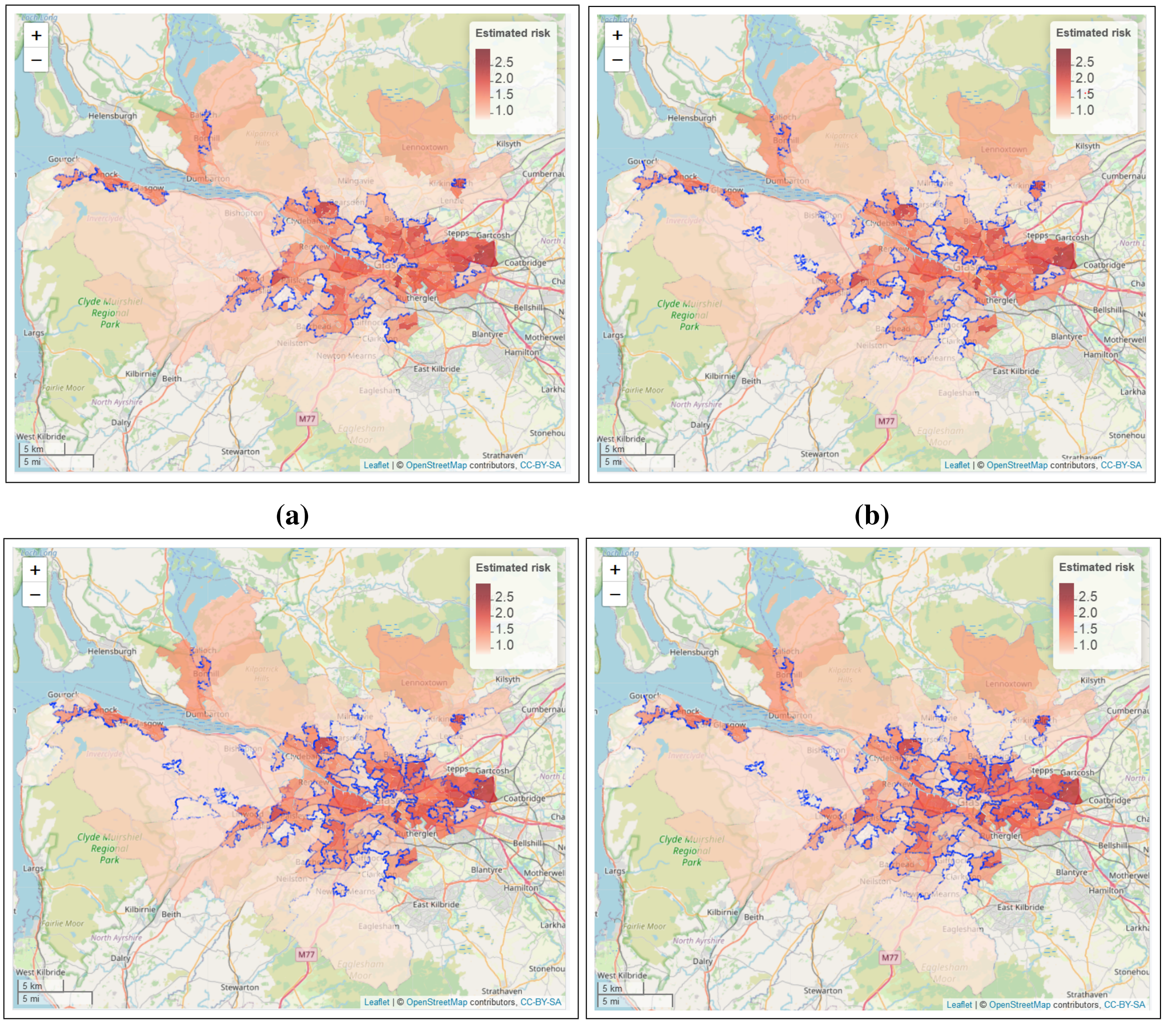
5.6.4.2 Posterior uncertainty in the estimated cluster structure

To illustrate the uncertainty in the estimated cluster structure, Figure 5.5 summarises the posterior distribution of $\widetilde{\boldsymbol{W}}$ obtained from the ten Markov chains for model **ST-A***, which assumes a single cluster structure for all time periods. In the figure each grid square represents a candidate neighbourhood matrix $\boldsymbol{W}^{(c,k)}$ corresponding to a distinct cluster structure, where the horizontal axis denotes the number of clusters and the vertical axis denotes the clustering method. The grid square on the bottom left corner corresponds to the border sharing $\boldsymbol{W} = \boldsymbol{W}^{(1,1)}, \ldots, \boldsymbol{W}^{(M,1)}$ (i.e. $k = 1$) which represents no clusters in disease risk. The figure shows that this no cluster solution is not supported by the data, with a posterior probability of zero. The figure also shows that the posterior distribution is mainly centered on 8 different cluster structures, which each have posterior probabilities above 0.06. The top four of these cluster structures have posterior probabilities of 0.3, 0.138, 0.1 and 0.1 respectively, and are displayed in Figure 5.6 and denoted by (a), (b), (c) and (d). Cluster structure (a), which has the highest posterior probability, identifies 2 spatial clusters (risk levels), although from the map it is clear that this corresponds to many more spatially distinct clusters. In contrast, structures (b) to (d) identify 4 or 5 distinct cluster levels, which is why there are more spatially distinct clusters identified in panels (b) to (d) in the figure. The adjusted Rand Index values between these four cluster structures range between 0.49 and 0.77, suggesting moderate agreement between them. The figure shows that all four cluster structures appear to mostly correspond to sizeable changes in disease risk between adjacent IZs, suggesting that the clustering model can identify such spatially distinct clusters.



	1	2	3	4	5	6	7	8	9	10
kmedoids		0.3	0	0	0	0	0	0	0	0
kmeans		0	0	0.1	0	0	0	0	0	0
EM		0	0	0	0	0	0	0	0	0
div		0	0	0	0	0	0	0	0	0
agg_ward		0.001	0	0	0	0	0	0	0	0
agg_complete		0.099	0	0.004	0.096	0	0	0	0	0
agg_centroid		0	0.011	0.088	0.001	0	0	0	0	0
agg_average		0	0.062	0.138	0.1	0	0	0	0	0
	0									

Figure 5.5: Summary of the posterior distribution of $\widetilde{\boldsymbol{W}}$ over 10 Markov chains from model **ST-A***.



(a)

(b)

(c)

(d)

Figure 5.6: The four most likely cluster structures selected by model **ST-A***, which are represented by blue dots. The colour shading for the areas denotes posterior median disease risk in 2014 (the middle of the study period).

5.6.5 Computational time

Table 5.7 displays the time taken to fit each of the five models to the motivating Greater Glasgow and Clyde Health Board respiratory disease data. The run times relate to a single Markov chain containing 100,000 samples with a burn-in period of 80,000, which is then thinned by 10. All models are run on an HP computer with an Intel Core i7-7700 CPU 3.60 GHz processor and 16GB of RAM. The table shows that model **ST-N** is the fastest of the five models, which is because it uses the border sharing $\boldsymbol{W}$ and so does not need to estimate the neighbourhood matrix within the modelling process as the other models do. However, the clustering models only have to be fitted once to the data to estimate the cluster structure. In contrast, if model **ST-N** was fitted separately with each candidate cluster structure generated in stage one of our approach, and then the best structure was chosen via a model comparison rule, then it would have to be fitted around 70 times. Thus using model **ST-N** in this fashion would be much computationally slower than using any of the cluster models proposed here. When comparing the speed of the clustering models, the table shows that models **ST-B** and **ST-B*** are slower than **ST-A** and **ST-A***, which is because they need to estimate a separate neighbourhood matrix for each time period. Additionally, models **ST-A*** and **ST-B*** that estimate the spatial dependence parameters (ρ_s, ρ_{s_t}) from the data are naturally slower than models **ST-A** and **ST-B** that treat these parameters as fixed.

Table 5.7: Comparison of the computational time required to apply each model to the motivating data.

Model	Inference	Elapsed Time
ST-A	MCMC (with C++)	826.31s
ST-A*	MCMC (with C++)	866.00s
ST-B	MCMC (with C++)	1358.05s
ST-B*	MCMC (with C++)	2345.14s
ST-N	MCMC (with C++)	169.52s

5.7 Discussion

Smoothing models based on geographical adjacency are commonly used to estimate risk in disease mapping studies, and they force geographical neighbours to have similar disease risks. However, this smoothing will mask any discontinuities present in the risk surface, leading to sub-optimal risk estimation and no cluster identification. The latter is important because health agencies often target additional resources at communities with the greatest need, thus they need to identify the spatial extent of a cluster of high-risk areas. This chapter has proposed a novel clustering-based adjacency modelling approach in the spatio-temporal domain, which can simultaneously estimate disease risk and identify the locations of clusters of high/low risk areas that may be static or evolve dynamically over time. The methodology first constructs a large collection of candidate cluster structures for the data, which each corresponds to a candidate neighbourhood matrix. Then a spatio-temporal model is fitted to the data that jointly estimates disease risk and the cluster structure, the latter by treating the neighbourhood matrix as a parameter to be estimated from the set of candidate structures constructed in stage 1. As this matrix determines the spatial correlation structure in the data, the approach extends the standard practice in areal unit modelling that naively assumes the border sharing neighbourhood matrix provides a suitable spatial correlation structure for the data. In fact, the methodology parallels the standard practice in geostatistics, where an appropriate spatial dependence structure is identified for the data (e.g. by variogram analysis) rather than a single structure being assumed without assessing its suitability.

Existing cluster based methods (Anderson et al., 2016, Adin et al., 2019) as well as the approach introduced in Chapter 3 fit a separate model to each candidate cluster structure, and identify the optimal cluster structure by model comparison approaches. In contrast, the approach here has the advantages of having substantially reduced computational time (it only fits a single spatio-temporal model), and allowing the uncertainty in the cluster structure to be quantified when estimating disease risk in the second step. For example, although the computational time of model **ST-A*** is almost 60% longer than that of the same model with a fixed neighbourhood matrix (or cluster structure), the latter would have to be fitted 73 times in a model comparison setting because we have 73 candidate cluster structures to consider. One approach that does allow for cluster uncertainty is Anderson et al. (2016) and I have extended this approach by considering a spatio-temporal rather than a spatial domain and proposing cluster models where the spatial clusters either remain fixed or evolve dynamically over time. The simulation study shows that the proposed models provide accurate risk estimates in the presence of clusters (discontinuities), particularly performing better than a similar non-cluster model. This improved performance is because our models account for the clusters in risk by estimating an appropriate neighbourhood matrix, which better represents the spatial autocorrelation structure in the data, therefore removing any redundant smoothing of the spatial random effects between neighbours. The study also shows that our models can accurately identify both static and temporally dynamic clusters, with high ARI values being obtained in both cases. However, as expected cluster identification is more accurate for static rather than dynamic clusters, which is due to the former having more data with which to identify the high-risk clusters due to them recurring for multiple time periods. Additionally, the simulation results suggest that our models are less accurate at estimating disease risk and identifying the correct clusters when the number of expected cases in each areal unit is very small (SF = 4), because there are not sufficient data from which to identify more prominent clusters. Therefore, I recommend using the proposed models for non-rare diseases with moderate to large numbers of expected cases, e.g. greater than 40 expected cases in each areal unit.

Here a range of classical clustering techniques are used to identify the candidate cluster structures rather than scan statistics (Kulldorff, 1997, Takahashi et al., 2008, Kulldorff et al., 2009) because the latter only identify a relatively small number of clusters of areas exhibiting high-risks rather than partitioning the risk surface into different risk levels, which is required here for constructing candidate neighbourhood matrices. In our method the maximum number of clusters K denotes the maximum number of risk levels and not the maximum number of spatially distinct clusters, which is illustrated in Figure 5.4. In this chapter I have chosen $K = 10$, which has been shown to be a conservative (overly large) choice because the posterior distribution in Figure 5.5 has no posterior mass above 5 clusters. Note that the choice of K does not depend on the number of areas, because it does not relate in any way to the maximum number of spatially connected clusters that can be identified. Thus the factors to consider when choosing K is that if K is too small then the true cluster structure may not exist in the candidates, whereas if K is too large then the computational cost increases because longer MCMC runs are likely to be needed for convergence with such a large number of candidates.

The motivating application also illustrates the superiority of the clustering-based models compared to non-clustering alternatives, with the proposed models able to produce a better model fit to the data and provide additional insight as to the locations of high-risk clusters. In regard to the latter, the majority of the identified cluster discontinuities occur between geographical neighbours that exhibit very different disease risks, which will allow health agencies to better identify these high-risk areas and target additional resources where they are most needed. Each of the cluster models has its own appealing features, and the choice between them will depend on the aim of the analysis. The model with constant clusters over time may be more appropriate if the disease data have a high correlation in time and the main aim is to identify overall clusters/discontinuities for the entire study period. In contrast, if the disease data are less correlated in time and the cluster structures in particular years are of interest, then the model with temporally evolving clusters is likely to be the better choice. However, the latter model can sometimes pick out apparently erroneous clusters, due to the presence of random year to year fluctuations in the observed disease counts. Therefore in future work I will investigate a hybrid approach of the two considered here based on a $2q+1$ years moving window, where the candidate cluster structures for a given year constructed in stage 1 are obtained by clustering the data for the year in question and the q years before and after. Additional potential extensions of the approach developed here include adapting it for use with different spatio-temporal random effects structures, such as those of Knorr-Held (2000) and Rushworth et al. (2014) as well as utilising it in the context of a spatio-temporal multivariate disease model, which allows for simultaneously estimating the risk of multiple diseases in space and time. So far, the methodology that has been proposed in this thesis allows for discontinuities in the spatial risk pattern, by identifying clusters of areas that exhibit substantially different risks compared to their neighbours. In Chapter 6, I will introduce a boundary analysis approach to allow for such discontinuities in disease risk.

Chapter 6

Estimating spatio-temporal disease risks via a boundary analysis approach

6.1 Introduction

Research on detecting the spatial discontinuities in disease risk includes the fields of spatial clustering (Knorr-Held and Raßer, 2000, Anderson et al., 2014, Adin et al., 2019) and boundary analysis (Lu et al., 2007, Lee and Mitchell, 2013, Lee et al., 2021). The previous chapters focus on spatial clustering and develop approaches that allow for spatial discontinuities in the disease risk surface by partitioning the study region into disjoint clusters of areas with elevated or reduced risks compared to their geographical neighbours. In Chapter 3, the method identifies spatial clusters by eliciting a set of candidate cluster configurations using k-means clustering and then fitting separate Bayesian hierarchical models to all configurations. The most appropriate cluster structure is chosen by model comparison techniques. In Chapter 4, I propose a Bayesian spatial model with the optimal cluster structure estimated as a parameter during the modelling process, which allows us to quantify the uncertainty in the cluster structure. The model has been extended to the spatio-temporal domain in Chapter 5, where the spatial clusters either remain fixed or evolve dynamically over time. These approaches produce closed boundaries which enclose an area or groups of areas that have very different risks from their neighbours. By contrast, boundary analysis, rather than looking for spatial clusters, aims to find the locations where geographically adjacent areas have very different disease risks. These locations correspond to "boundaries" (large or small) in the risk surface, which can be open and do not necessarily completely enclose an area or group of units. The majority of the boundary analysis approaches treat each w_{ij} element in the neighbourhood matrix as a binary random quantity if areas (i, j) share a common geographical border, rather than assuming it is fixed at 1. If $w_{ij} = 0$ then a boundary is said to exist between the two adjacent areas in the risk surface, while $w_{ij} = 1$ corresponds to no risk boundary. One of the most recent such approaches is proposed by Lee et al. (2021), who estimate an appropriate neighbourhood matrix for the data using a graph-based optimisation algorithm, and then fit a Bayesian spatio-temporal model based on this estimated matrix. However, a potential limitation to this approach is that since the optimisation algorithm only automatically provides a single neighbourhood matrix in the first step, it does not take the uncertainty associated with the neighbourhood matrix into account when estimating disease risk. To overcome this issue, in this chapter I extend the approach of Lee et al. (2021) by obtaining multiple candidate neighbourhood matrices via the graph-based optimisation algorithm, and then allowing for variation in the neighbourhood matrix, and hence in the boundaries identified in the modelling procedure. In addition, model inference is based on a Bayesian setting via a Metropolis-coupled Markov chain Monte Carlo algorithm due to the multi-modal posterior distributions.

The methodology is motivated by a study of respiratory disease risk in the Greater Glasgow and Clyde Health Board during the time period from 2011 to 2017. The remainder of this chapter is organised as follows. Section 6.2 discusses the background spatio-temporal risk model. Section 6.3 presents the proposed methodology, which identifies boundaries in the risk surface via estimation of the neighbourhood matrix. The efficacy of this approach is evidenced using simulations in Section 6.4, while the sensitivity of the approach to the choice of hyperpriors is examined in Section 6.5. Section 6.6 applies the methodology to the motivating data introduced in Section 5.2.1, while Section 6.7 summarises the main findings in this chapter and discusses the ideas for future work.

6.2 Spatio-temporal modelling of areal unit count data

The observed and expected disease counts for areal unit $i = 1, \ldots, n$ and time period $t = 1, \ldots, T$ are denoted by $\{Y_{it}\}$ and $\{E_{it}\}$ respectively. Covariate information (if relevant) is given by $\{\boldsymbol{x}_{it}\}$, where $\boldsymbol{x}_{it}^{\top} = (1, x_{it1}, \ldots, x_{itp})$ contains a vector of p known covariates relating to areal unit i during time period t and a 1 for the intercept term. As the response variable is a count, the most commonly used data likelihood model is given by $Y_{it}|E_{it}, R_{it} \sim \text{Poisson}(E_{it}R_{it})$, where R_{it} represents the overall disease risk in areal unit i during time period t relative to the expected count E_{it}, and is on the same scale as the SIR. A general Bayesian hierarchical model commonly specified for these data is given by

$$\begin{aligned} Y_{it}|E_{it},R_{it} &\sim \text{Poisson}(E_{it}R_{it}),\ i=1,\ldots,n;\quad t=1,\ldots,T, \\ \ln(R_{it}) &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it}, \\ \beta_j &\sim \text{N}(0,1000),\ \text{for } j=0,\ldots,p. \end{aligned} \tag{6.1}$$

The spatio-temporal pattern in disease risk is modelled by known covariates $\{\boldsymbol{x}_{it}\}$ with regression parameters $\boldsymbol{\beta}=(\beta_0,\ldots,\beta_p)$, and random effects $\{\phi_{it}\}$. The latter are included in the model to account for any residual spatio-temporal autocorrelation after adjusting for covariates. Here I utilise the spatio-temporal structure proposed by Rushworth et al. (2014), which has the autoregressive decomposition

$$\begin{aligned} \boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} &\sim \text{N}\left(\alpha\boldsymbol{\phi}_{t-1},\tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1}\right),\ t=2,\ldots,T, \\ \boldsymbol{\phi}_1 &\sim \text{N}\left(\mathbf{0},\tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1}\right), \\ \rho,\alpha &\sim \text{Uniform}(0,1), \\ \tau^2 &\sim \text{Inverse-Gamma}(1,0.01), \end{aligned} \tag{6.2}$$

where $\boldsymbol{\phi}_t=(\phi_{1t},\ldots,\phi_{nt})$ represents the spatial surface for time period t. Temporal autocorrelation is induced amongst $\boldsymbol{\phi}_t$ via a multivariate first order autoregressive process. $\alpha\in[0,1]$ is the temporal dependence parameter, with a value of 0 corresponding to temporal independence while a value of 1 indicates strong temporal autocorrelation. The random effects at time point 1, $\boldsymbol{\phi}_1=(\phi_{11},\ldots,\phi_{n1})$, are specified using the Leroux CAR prior (Leroux et al., 2000), thus the spatial autocorrelation is induced through the precision matrix $\boldsymbol{Q}(\rho,\boldsymbol{W})=\rho(\text{diag}(\boldsymbol{W}\mathbf{1})-\boldsymbol{W})+(1-\rho)\boldsymbol{I}$, where $\mathbf{1}$ is an $n\times 1$ vector of ones and $\boldsymbol{I}$ is an $n\times n$ identity matrix. In Rushworth et al. (2014), the spatial autocorrelation structure in the data is fixed and represented by the $n\times n$ border sharing neighbourhood matrix $\boldsymbol{W}$, where $w_{ij}=1$ if areas (i,j) share a common geographical border (denoted $i\sim j$) and is 0 otherwise (diagonal elements $w_{ii}=0$). The univariate full conditional distribution corresponding to the Leroux prior for area i at time point one is given by

$$\phi_{i1}|\boldsymbol{\phi}_{-i1}\sim\text{N}\left(\frac{\rho\sum_{j=1}^n w_{ij}\phi_{j1}}{\rho\sum_{j=1}^n w_{ij}+1-\rho},\frac{\tau^2}{\rho\sum_{j=1}^n w_{ij}+1-\rho}\right), \tag{6.3}$$

where $\boldsymbol{\phi}_{-i1}=(\phi_{11},\ldots,\phi_{i-1,1},\phi_{i+1,1},\ldots,\phi_{n1})$. Here ρ is the spatial dependence parameter, with a value of 1 corresponding to strong spatial autocorrelation whereas a value of 0 corresponds to independence in space (as $\phi_{i1} \sim \mathrm{N}(0, \tau^2)$). However, since this model uses the border sharing $\boldsymbol{W}$ to represent the spatial correlation structure, it enforces spatial autocorrelation between all pairs of geographically neighbouring areas, which may lead to over-smoothing of the estimated disease risk maps and hinder the detection of boundaries in the risk surfaces. Therefore in the next section I utilise the above spatio-temporal structure to model the data, with the difference that the neighbourhood matrix $\boldsymbol{W}$ is treated as a parameter to be estimated from the data, rather than being determined simply based on geographical adjacency.

6.3 Methodology

I propose a two-stage extension of the approach proposed by Lee et al. (2021), which can jointly estimate disease risk and identify the locations of boundaries in the spatial surface that separate two geographically adjacent areas exhibiting very different risks. The approach consists of two stages. In stage 1, the graph-based optimisation algorithm proposed by Lee et al. (2021) is applied to the data M times to obtain M candidate neighbourhood matrices that allow for spatial boundaries in the data. In stage 2, a spatio-temporal model is fitted to the data, in which the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ is treated as a parameter to be estimated from the set of candidates constructed in stage 1. This stage allows for uncertainty in $\widetilde{\boldsymbol{W}}$ and hence the boundaries, which is our main methodological contribution compared to Lee et al. (2021).

6.3.1 Stage 1 — Generating neighbourhood matrices that account for boundaries in disease risk

6.3.1.1 Estimating the residual spatial risk surface

The random effects $\boldsymbol{\phi}_t$ model the residual spatial variation in the data at time period t after adjusting for covariates. By our approach, boundaries are incorporated into the model via the neighbourhood matrix, which means that they relate to the random effects $\{\phi_{it}\}$ (see model (6.2)). Hence the fist step is to estimate $\{\phi_{it}\}$ from the data and the general model (6.1) by

$$\tilde{\phi}_{it} = \ln\left(\frac{\mathbb{E}(Y_{it})}{E_{it}}\right) - \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} \approx \ln\left(\frac{Y_{it}}{E_{it}}\right) - \boldsymbol{x}_{it}^{\top}\hat{\boldsymbol{\beta}}. \tag{6.4}$$

As in Chapter 5, the above approximation replaces the unknown $\mathbb{E}(Y_{it})$ with the observed data Y_{it}, The regression parameters are estimated for this initial stage by assuming independence via maximum likelihood estimation, and are denoted by $\hat{\boldsymbol{\beta}}$. Since the residual spatial surfaces from the motivating data are similar over time, with an average Pearson's correlation coefficient of 0.84 between each pair of years, here I only consider the scenario of constant risk boundaries over time. In this case, a single residual spatial risk surface is estimated by averaging the random effects for each areal unit over the T time periods, which is

$$\tilde{\phi}_i = \frac{1}{T}\sum_{t=1}^{T}\tilde{\phi}_{it}, \text{ for } i = 1,\ldots,n. \tag{6.5}$$

This average residual spatial surface $\tilde{\boldsymbol{\phi}} = \{\tilde{\phi}_i\}$ is used in the graph-based optimisation algorithm described in the next section to estimate candidate neighbourhood matrices.

6.3.1.2 Graph-based optimisation

Lee et al. (2021) proposed a graph-based optimisation algorithm for estimating an appropriate neighbourhood matrix for the data. The algorithm views the entire study region as a graph $\mathcal{G}$, whose vertex-set $\mathcal{V}(\mathcal{G})$ is the set of n areal units that comprise the study region, and whose edge-set $\mathcal{E}(\mathcal{G})$ is defined by the border sharing $\boldsymbol{W} = \{w_{ij}\}$ of the graph $\mathcal{G}$ via $\mathcal{E}(\mathcal{G}) = \{(i,j)|w_{ij} = 1\}$. The goal of the algorithm is to estimate which edges in $\mathcal{E}(\mathcal{G})$ should be removed (or retained) based on an objective function. Suppose $\tilde{\mathcal{G}}$ is a subgraph of $\mathcal{G}$ (i.e. a graph $\tilde{\mathcal{G}}$ whose vertex set and edge set are subsets of those of $\mathcal{G}$), and the neighbourhood matrix corresponding to this subgraph $\tilde{\mathcal{G}}$ is denoted by $\boldsymbol{W}_{\tilde{\mathcal{G}}} = \{w_{\tilde{\mathcal{G}}_{ij}}\}$. Given the input $\tilde{\boldsymbol{\phi}}$ and the border sharing $\boldsymbol{W}$, the algorithm estimates an optimised neighbourhood matrix $\boldsymbol{W}_{\tilde{\mathcal{G}}}$ by finding a suitable spanning subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$ (i.e. $\tilde{\mathcal{G}}$ is a subgraph of $\mathcal{G}$ and they also have the same vertex set) that maximises the value of an objective function $S(\tilde{\boldsymbol{\phi}})$. If the edge between vertices/areas (i,j) is removed in the optimal graph $\tilde{\mathcal{G}}$ then we set $w_{\tilde{\mathcal{G}}_{ij}} = 0$, which suggests that a boundary exists between areas i and j as the random effects between them are conditionally independent in space given $w_{\tilde{\mathcal{G}}_{ij}} = 0$ (as specified by the objective function, see below). If the edge is retained in $\tilde{\mathcal{G}}$ then we have $w_{\tilde{\mathcal{G}}_{ij}} = 1$, suggesting no boundary between areas (i,j) as their random effects are conditionally correlated. The objective function is built on the natural log of the product of full conditional distributions $f(\tilde{\phi}_i|\tilde{\boldsymbol{\phi}}_{-i})$ from the intrinsic CAR prior (Besag et al., 1991), because this is a commonly used spatial correlation model that uses the neighbourhood matrix to determine the correlation structure. The objective function with respect to a spanning subgraph $\mathcal{H}$ of $\mathcal{G}$ with the corresponding neighbourhood matrix $\boldsymbol{W}_{\mathcal{H}} = \{w_{\mathcal{H}_{ij}}\}$ is given by

$$\begin{aligned}
S(\tilde{\boldsymbol{\phi}},\mathcal{H}) &= \ln\left[\prod_{i=1}^{n} f(\tilde{\phi}_i|\tilde{\boldsymbol{\phi}}_{-i})\right] \\
&= \ln\left[\prod_{i=1}^{n} \mathrm{N}\left(\frac{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\tilde{\phi}_j}{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}}\right)\right] \\
&\propto -\frac{n}{2}\ln(\tau^2) + \frac{1}{2}\sum_{i=1}^{n}\ln\left(\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\right) - \\
&\quad \frac{1}{2\tau^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\right)\left(\tilde{\phi}_i - \frac{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\tilde{\phi}_j}{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}}\right)^2 .
\end{aligned} \tag{6.6}$$

The unknown variance parameter τ^2 is estimated by maximising $S(\tilde{\boldsymbol{\phi}},\mathcal{H})$ with respect to τ^2, which gives $\hat{\tau}^2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\right)\left(\tilde{\phi}_i - \frac{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\tilde{\phi}_j}{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}}\right)^2 \Big/ n$. This variance estimator $\hat{\tau}^2$ is plugged into equation (6.6) to produce the final objective function, which is

$$S(\tilde{\boldsymbol{\phi}},\mathcal{H}) \propto \frac{1}{2}\sum_{i=1}^{n}\ln\left(\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\right) - \frac{n}{2}\ln\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\right)\left(\tilde{\phi}_i - \frac{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}\tilde{\phi}_j}{\sum_{j=1}^{n} w_{\mathcal{H}_{ij}}}\right)^2\right]. \tag{6.7}$$

To reduce the computational burden, the algorithm operates via an iterative local search method, in which the vertices of the original graph $\mathcal{G}$ are considered in some fixed order. At the first step the algorithm considers the first vertex i (i.e. area i) and uses the original graph $\mathcal{G}$ to decide whether each of the edges linked to i (i.e. each neighbour relation of area i) should be removed or not. If deleting an edge can increase the objective function then it should be deleted from the graph $\mathcal{G}$. Based on the edges that have been deleted a new subgraph $\mathcal{G}^{'}$ is obtained. Then the algorithm continues with the next vertex and considers the objective function with respect to $\mathcal{G}^{'}$ this time. It continues in this way, returning to the first vertex when reaching the last vertex, until all remaining feasible vertices are passed through without identifying any deletions that increase the objective function. The algorithm ensures every vertex in $\mathcal{V}(\mathcal{G})$ must retain at least one edge linked to it. For more details on the graph-based optimisation algorithm, see Lee et al. (2021). This optimisation algorithm is implemented using the spatio-temporal modelling package CARBayesST (Lee et al., 2018) in R (R Core Team, 2013).

In order to quantify the uncertainty in the neighbourhood matrix when modelling spatio-temporal areal unit count data, I apply the above graph-based optimisation algorithm M times to the average residual spatial surface $\tilde{\boldsymbol{\phi}} = \{\tilde{\phi}_i\}$ estimated from equation (6.5), with each time considering the vertices of the graph $\mathcal{G}$ in a different order in the algorithm. Finally, this leads to M candidate neighbourhood matrices, denoted by $\left(\boldsymbol{W}_G^1, \boldsymbol{W}_G^2, \ldots, \boldsymbol{W}_G^M\right)$. These candidate matrices correspond to a range of possible boundaries in the data, where some edges are present and the corresponding random effects are smoothed towards each other, whereas other edges are removed so smoothing is not enforced between the corresponding neighbouring random effects.

The value of M determines the size of the sample space for $\widetilde{\boldsymbol{W}}$ in the proposed model (6.8). As each candidate in the set $\left(\boldsymbol{W}_G^1, \boldsymbol{W}_G^2, \ldots, \boldsymbol{W}_G^M\right)$ corresponds to a different ordering of the vertices applied in the algorithm, and the motivating data have a total of 257 vertices (IZs), theoretically the maximum number of candidate matrices is up to 257!, which is infeasible to obtain computationally. Here I choose $M = 100$ and the 100 different orderings of the vertices are obtained by randomly sampling 100 permutations from the 257! possible permutations. The graph-based optimisation algorithm takes around two hours to obtain the 100 candidate neighbourhood matrices for the Glasgow disease data analysed in Section 6.6.

6.3.2 Stage 2 — Bayesian spatio-temporal modelling

The second stage of the approach fits a model to the data that simultaneously estimates the spatio-temporal trend in disease risk and identifies the boundaries in the risk surface. The proposed model is given by

$$
\begin{aligned}
Y_{it}|E_{it}, R_{it} &\sim \text{Poisson}(E_{it}R_{it}) \;\; i = 1, \ldots, n; \;\; t = 1, \ldots, T, \\
\ln(R_{it}) &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{it}, \qquad (6.8) \\
\beta_j &\sim \text{N}(0, 1000), \text{ for } j = 0, \ldots, p, \\
\boldsymbol{\phi}_1 &\sim \text{N}\left(\boldsymbol{0}, \tau^2 \tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right), \\
\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} &\sim \text{N}\left(\alpha\boldsymbol{\phi}_{t-1}, \tau^2 \tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right), \;\; t = 2, \ldots, T, \\
\widetilde{\boldsymbol{W}} &\sim \text{Discrete Uniform}\left(\boldsymbol{W}_G^1, \ldots, \boldsymbol{W}_G^M\right), \\
\alpha &\sim \text{Uniform}(0, 1), \\
\tau^2 &\sim \text{Inverse-Gamma}(1, 0.01).
\end{aligned}
$$

This model differs from the model proposed by Rushworth et al. (2014) (see Section 6.2) in two main ways. Firstly, here I treat the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ as a parameter, and assign it a discrete uniform prior whose values are the set of candidate neighbourhood matrices $(\boldsymbol{W}_G^1, \ldots, \boldsymbol{W}_G^M)$ which are previously constructed via the graph-based optimisation algorithm, rather than naively using the border sharing $\boldsymbol{W}$. Secondly, the Leroux CAR prior in the model of Rushworth et al. (2014) is replaced by the intrinsic CAR prior (where $\rho = 1$) in the proposed model. This is because our approach models the spatial autocorrelation structure locally for each pair of neighbouring areas by estimating $\widetilde{\boldsymbol{W}}$ for the data, which may make the estimation of a single global spatial dependence parameter redundant. The intrinsic CAR model enforces strong spatial autocorrelation globally, so that the spatial correlation structure can be adjusted locally by estimating $\widetilde{\boldsymbol{W}}$, rather than globally by a dependence parameter ρ. The precision matrix corresponding to the intrinsic CAR prior is $\boldsymbol{Q}(\widetilde{\boldsymbol{W}}) = \text{diag}(\widetilde{\boldsymbol{W}}\mathbf{1}) - \widetilde{\boldsymbol{W}}$, which is singular. However, the invertibility of the precision matrix is required because its determinant needs to be calculated when updating the parameter $\widetilde{\boldsymbol{W}}$. Therefore, in order to ensure the precision matrix is diagonally dominant and hence invertible, the singular precision matrix $\boldsymbol{Q}(\widetilde{\boldsymbol{W}})$ is replaced by an invertible precision matrix $\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}}) = \text{diag}(\widetilde{\boldsymbol{W}}\mathbf{1}) - \widetilde{\boldsymbol{W}} + \varepsilon\boldsymbol{I}$ (Lee et al., 2014), which adds a small positive constant ε onto the diagonal terms of $\boldsymbol{Q}(\widetilde{\boldsymbol{W}})$. Rushworth et al. (2017) have shown that a small value of ε ($\varepsilon < 0.01$) does not affect estimation results, hence I set $\varepsilon = 0.00001$ when implementing the model. The random effects $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T\}$ are modelled by a spatially autocorrelated multivariate first order autoregessive process, where temporal autocorrelation is modelled via the mean $\alpha\boldsymbol{\phi}_{t-1}$, and spatial autocorrelation in the data is modelled by the precision matrix $\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})$. Finally, each regression parameter β_j is assigned a Gaussian prior distribution with mean zero and variance 1000. The temporal autocorrelation parameter $\alpha \in [0,1]$ and the variance parameter τ^2 are assigned a weakly informative uniform prior, $\alpha \sim \text{Uniform}(0,1)$, and an Inverse-Gamma prior, $\tau^2 \sim \text{Inverse-Gamma}(1, 0.01)$. To achieve identifiability, the random effects are zero-mean centred. As in previous chapters, the model is outlined in its most general form that includes covariate information, but covariates are not included in the application study in Section 6.6, in other words, $\boldsymbol{x}_{it}^\top \boldsymbol{\beta} = \beta_0$. This is because the aim of the analysis is to identify spatial boundaries in the disease risk surface, rather than in the residual risk surface after adjusting for covariate factors.

6.3.3 Inference

As in the previous chapters, I initially carried out model inference in a Bayesian setting via Markov chain Monte Carlo (MCMC) simulation, using both the Gibbs sampling (Geman and Geman, 1984) and Metropolis-Hastings steps (Metropolis et al., 1953, Hastings, 1970). However, initial simulations showed that standard implementations of MCMC simulation lead to poor mixing of $\widetilde{\boldsymbol{W}}$ in the Markov chain, which is because the posterior probability distribution of $\widetilde{\boldsymbol{W}}$ contains multiple modes. These modes represent high probability regions, and when they are separated by low probability regions, a Markov chain currently exploring a peak of high probability may experience difficulty crossing the low probability regions to explore other peaks (Altekar et al., 2004). As a result, the updates of $\widetilde{\boldsymbol{W}}$ often get trapped in a local mode. To address this problem, I adopt the Metropolis-coupled Markov chain Monte Carlo ($(\mathrm{MC})^3$) algorithm used by Napier et al. (2019). The $(\mathrm{MC})^3$ algorithm runs multiple Markov chains in parallel at different temperature levels and then couples the chains together to prevent them from becoming stuck in a local mode. A higher temperature level makes a chain accept more proposed moves, thus allowing it to more readily jump between multiple modes in the posterior distribution.

6.3.3.1 $(\mathrm{MC})^3$ algorithm

Suppose the $(\mathrm{MC})^3$ algorithm runs V Markov chains in parallel, where each chain is labeled by $v \in (1,2,\ldots,V)$. The temperature level for chain v is denoted by T_v, and we have $0 < T_V < T_{V-1} <,\ldots,< T_2 < T_1 = 1$. The first chain with $T_1 = 1$ is also known as the cold chain, and the posterior samples from the cold chain are used for model inference. $\boldsymbol{\Omega}_{vl}$ denotes the collection of model parameters at the l_{th} iteration of the Markov chain v and in our context $\boldsymbol{\Omega}_{vl} = (\boldsymbol{\beta}_{vl}, \boldsymbol{\phi}_{vl}, \widetilde{\boldsymbol{W}}_{vl}, \tau^2_{vl}, \alpha_{vl})$. The $(\mathrm{MC})^3$ algorithm is presented as follows.

1. Set starting values $\boldsymbol{\Omega}_{v0} = (\boldsymbol{\beta}_{v0}, \boldsymbol{\phi}_{v0}, \widetilde{\boldsymbol{W}}_{v0}, \tau^2_{v0}, \alpha_{v0})$ in each chain for $v = 1,2,\ldots,V$.

2. Repeat the following steps for each sampling iteration $l = 1,2,\ldots,L$.

   (a) At iteration l repeat the following steps for each Markov chain for $v = 1,2,\ldots,V$, and each model parameter $\omega_{vl} \in \boldsymbol{\Omega}_{vl}$.

      i. Propose a new value for ω_{vl}, called ω^*_{vl}, from a proposal distribution $g(\omega^*_{vl}|\omega_{vl})$.

ii. Accept ω^*_{vl} with probability p_1,

$$p_1 = \min\left\{ \frac{f(\omega^*_{vl}|\boldsymbol{Y})^{T_v}/g(\omega^*_{vl}|\omega_{vl})}{f(\omega_{vl}|\boldsymbol{Y})^{T_v}/g(\omega_{vl}|\omega^*_{vl})}, 1 \right\},$$

where $f(\cdot)$ represents the full conditional distribution of ω_{vl} or ω^*_{vl}.

iii. Generate a random variable, U_1, that is uniformly distributed on the interval $[0,1]$; If $U_1 \leq p_1$, accept ω^*_{vl} as the next value in the chain v, i.e. $\omega_{v,l+1} = \omega^*_{vl}$. Otherwise, $\omega_{v,l+1} = \omega_{vl}$.

(b) Randomly select two of the chains to couple the chains, e.g. chains j and k, and exchange their values.

i. Swap chains j and k with probability p_2, where

$$p_2 = \min\left\{ \frac{f(\boldsymbol{\Omega}_{kl}|\boldsymbol{Y})^{T_j} f(\boldsymbol{\Omega}_{jl}|\boldsymbol{Y})^{T_k}}{f(\boldsymbol{\Omega}_{jl}|\boldsymbol{Y})^{T_j} f(\boldsymbol{\Omega}_{kl}|\boldsymbol{Y})^{T_k}}, 1 \right\}.$$

ii. Generate a uniform random sample $U_2 \sim \text{Uniform}(0,1)$; If $U_2 \leq p_2$, then the proposed swap is accepted and chains j and k exchange their values.

The $(\text{MC})^3$ algorithm is not applied to the parameters that are sampled using Gibbs sampling. The temperatures are determined by a geometric progression, which is a common choice in the literature (Kofke, 2002, Earl and Deem, 2005). The geometrically spaced temperatures are given by $T_{v+1} = c * T_v$, with a scale factor $c \in (0,1)$. The value of c is altered within the algorithm to ensure the swaps of two chains are accepted between 20 % and 30% of the time, thereby providing a sufficient amount of mixing (Napier et al., 2019). The number of chains needed for adequate mixing can depend on the complexity of the data (Altekar et al., 2004). In this chapter I run $V = 5$ coupled chains, which appears to result in good mixing for both simulated and real application data. The $(\text{MC})^3$ algorithm is written and implemented in R (R Core Team, 2013) and C++ via the R package Rcpp (Eddelbuettel et al., 2011, Eddelbuettel, 2013). The posterior distributions for each of the model parameters are described in the next section.

6.3.3.2 Posterior distributions for each parameter

Update $\boldsymbol{\beta}$

The full conditional distribution for $\boldsymbol{\beta}$ is

$$\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{Y}) &\propto \prod_{i=1}^{n}\prod_{t=1}^{T}\text{Poisson}(Y_{it}|\boldsymbol{\beta}) \times \prod_{j=0}^{p}\text{N}(\beta_j|0,\ 1000) \\
&\propto \left(\prod_{i=1}^{n}\prod_{t=1}^{T}\left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it})\right)^{Y_{it}}\right)\exp\left(-\sum_{i=1}^{n}\sum_{t=1}^{T}E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it})\right)\times\prod_{j=0}^{p}\exp\left(\frac{-\beta_j^2}{2000}\right),
\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)$ is drawn as a block for all p covariates, including an intercept term β_0.

Update ϕ_{it}

The joint distribution for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_T)$ can be written as

$$f(\boldsymbol{\phi}) = f(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_T) = f(\boldsymbol{\phi}_1)f(\boldsymbol{\phi}_2|\boldsymbol{\phi}_1)f(\boldsymbol{\phi}_3|\boldsymbol{\phi}_2,\boldsymbol{\phi}_1)\ldots f(\boldsymbol{\phi}_T|\boldsymbol{\phi}_{T-1},\ldots,\boldsymbol{\phi}_1).$$

Since $\boldsymbol{\phi}_1 \sim \text{N}\left(\boldsymbol{0},\tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right)$, $\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} \sim \text{N}\left(\alpha\boldsymbol{\phi}_{t-1},\tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right)$ for $t = 2,\ldots,T$, and denote $\tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}$ by $\boldsymbol{R}^{-1}$, we get

$$\begin{aligned}
f(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_T) &= f(\boldsymbol{\phi}_1)\prod_{t=2}^{T}f(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}) \\
&\propto \text{N}\left(\boldsymbol{\phi}_1|\boldsymbol{0},\boldsymbol{R}^{-1}\right)\prod_{t=2}^{T}\text{N}\left(\boldsymbol{\phi}_t|\alpha\boldsymbol{\phi}_{t-1},\boldsymbol{R}^{-1}\right) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\phi}_1^{\top}\boldsymbol{R}\boldsymbol{\phi}_1\right)\prod_{t=2}^{T}\exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}_t-\alpha\boldsymbol{\phi}_{t-1}\right)^{\top}\boldsymbol{R}\left(\boldsymbol{\phi}_t-\alpha\boldsymbol{\phi}_{t-1}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}_1^{\top}\boldsymbol{R}\boldsymbol{\phi}_1+\sum_{t=2}^{T}\left(\boldsymbol{\phi}_t^{\top}\boldsymbol{R}\boldsymbol{\phi}_t+\alpha^2\boldsymbol{\phi}_{t-1}^{\top}\boldsymbol{R}\boldsymbol{\phi}_{t-1}-2\alpha\boldsymbol{\phi}_t^{\top}\boldsymbol{R}\boldsymbol{\phi}_{t-1}\right)\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\phi}^{\top}\left(\boldsymbol{D}\otimes\boldsymbol{R}\right)\boldsymbol{\phi}\right),
\end{aligned}$$

where $\boldsymbol{D}$ is a $T \times T$ matrix, with each element D_{ts} given as

$$\boldsymbol{D}_{ts} = \begin{cases} 1+\alpha^2, & \text{if } t = s \neq T, \\ 1, & \text{if } t = s = T, \\ -\alpha, & \text{if } |t-s| = 1, \\ 0 & \text{if } |t-s| \geq 2. \end{cases}$$

Hence $\boldsymbol{\phi}$ follows a multivariate Gaussian distribution $\boldsymbol{\phi} \sim \mathrm{N}\left(\mathbf{0}, (\boldsymbol{D} \otimes \boldsymbol{R})^{-1}\right)$, where

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_T \end{bmatrix}_{nT \times 1} = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{n1} \\ \vdots \\ \phi_{1T} \\ \vdots \\ \phi_{nT} \end{bmatrix}_{nT \times 1} .$$

According to the conditional distribution property of a multivariate Gaussian distribution, the distribution of ϕ_{it} conditional on the remaining random effects $\boldsymbol{\phi}_{-it}$ is

$$\phi_{it} | \boldsymbol{\phi}_{-it} \sim \mathrm{N}\left(\mathrm{E}\left[\phi_t | \boldsymbol{\phi}_{-it}\right], \mathrm{Var}\left[\phi_{it} | \boldsymbol{\phi}_{-it}\right]\right),$$

where the conditional expectation is given by

$$\mathrm{E}[\phi_{it} | \boldsymbol{\phi}_{-it}] = \begin{cases} \dfrac{(1+\alpha^2) \sum_{j=1}^{n} w_{ij}\phi_{jt} - \alpha \sum_{j=1}^{n} w_{ij}\phi_{j,t+1} + \alpha \sum_{j=1}^{n} w_{ij}\phi_{i,t+1}}{(1+\alpha^2) \sum_{j=1}^{n} w_{ij}}, \text{ if } t = 1, \\ \dfrac{(1+\alpha^2) \sum_{j=1}^{n} w_{ij}\phi_{jt} - \alpha \sum_{j=1}^{n} w_{ij}\phi_{j,t-1} + \alpha \sum_{j=1}^{n} w_{ij}\phi_{i,t-1} - \alpha \sum_{j=1}^{n} w_{ij}\phi_{j,t+1} + \alpha \sum_{j=1}^{n} w_{ij}\phi_{i,t+1}}{(1+\alpha^2) \sum_{j=1}^{n} w_{ij}}, \text{if } t = 2, \ldots, T-1, \\ \dfrac{\sum_{j=1}^{n} w_{ij}\phi_{jt} - \alpha \sum_{j=1}^{n} w_{ij}\phi_{j,t-1} + \alpha \sum_{j=1}^{n} w_{ij}\phi_{i,t-1}}{\sum_{j=1}^{n} w_{ij}}, \text{ if } t = T. \end{cases}$$

The conditional variance is given by

$$\mathrm{Var}[\phi_{it} | \boldsymbol{\phi}_{-it}] = \begin{cases} \dfrac{\tau^2}{(1+\alpha^2) \sum_{j=1}^{n} w_{ij}}, & \text{if } t \neq T, \\ \dfrac{\tau^2}{\sum_{j=1}^{n} w_{ij}}, & \text{if } t = T. \end{cases}$$

Hence the full conditional distribution for ϕ_{it} can be computed by

$$\begin{aligned} f(\phi_{it}|Y_{it}) &\propto \text{Poisson}(Y_{it}|\phi_{it}) \times \text{N}(\phi_{it}|\boldsymbol{\phi}_{-it}) \\ &\propto \left(\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it})\right)^{Y_{it}} \exp\left(-E_{it}\exp(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}+\phi_{it})\right) \times \text{N}\left(\text{E}\left[\phi_t|\boldsymbol{\phi}_{-it}\right], \text{Var}\left[\phi_{it}|\boldsymbol{\phi}_{-it}\right]\right). \end{aligned}$$

Update α

The full conditional distribution for α is

$$\begin{aligned} f(\alpha|\boldsymbol{Y}) &\propto \prod_{t=2}^{T} \text{N}\left(\boldsymbol{\phi}_t|\alpha\boldsymbol{\phi}_{t-1}, \tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right) \times \text{Uniform}(0,1) \\ &\propto \text{N}(m,v), \end{aligned}$$

where

$$\begin{aligned} m &= \frac{\sum_{t=2}^{T} \boldsymbol{\phi}_t^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_{t-1}}{\sum_{t=2}^{T} \boldsymbol{\phi}_{t-1}^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_{t-1}}, \\ v &= \frac{\tau^2}{\sum_{t=2}^{T} \boldsymbol{\phi}_{t-1}^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_{t-1}}. \end{aligned}$$

Update τ^2

The full conditional distribution of τ^2 is given as

$$\begin{aligned} f(\tau^2|\boldsymbol{Y}) &\propto \text{N}\left(\boldsymbol{\phi}_1|\boldsymbol{0}, \tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right) \prod_{t=2}^{T} \text{N}\left(\boldsymbol{\phi}_t|\alpha\boldsymbol{\phi}_{t-1}, \tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right) \times \text{Inverse-Gamma}(1, 0.01) \\ &\sim \text{Inverse-Gamma}(\tilde{a}, \tilde{b}), \end{aligned}$$

where $\tilde{a} = 1 + \frac{nT}{2}$ and $\tilde{b} = 0.01 + \frac{1}{2}\left(\boldsymbol{\phi}_1^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_1 + \sum_{t=2}^{T}\left(\boldsymbol{\phi}_t - \alpha\boldsymbol{\phi}_{t-1}\right)^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\left(\boldsymbol{\phi}_t - \alpha\boldsymbol{\phi}_{t-1}\right)\right)$.

Update $\widetilde{\boldsymbol{W}}$

The full conditional distribution for $\widetilde{\boldsymbol{W}}$ is given as

$$\begin{aligned} f(\widetilde{\boldsymbol{W}}|\boldsymbol{Y}) &\propto \text{N}\left(\boldsymbol{\phi}_1|\boldsymbol{0}, \tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right) \prod_{t=2}^{T} \text{N}\left(\boldsymbol{\phi}_t|\alpha\boldsymbol{\phi}_{t-1}, \tau^2\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})^{-1}\right) \times f\left(\widetilde{\boldsymbol{W}} = \boldsymbol{W}_G^m\right) \\ &\propto ||\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})||^{\frac{T}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}_1^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\boldsymbol{\phi}_1 + \sum_{t=2}^{T}\left(\boldsymbol{\phi}_t - \alpha\boldsymbol{\phi}_{t-1}\right)^{\top}\tilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{W}})\left(\boldsymbol{\phi}_t - \alpha\boldsymbol{\phi}_{t-1}\right)\right)\tau^{-2}\right), \end{aligned}$$

where $||\cdot||$ denotes the determinant of a matrix.

In the model, $\widetilde{\boldsymbol{W}}$ is assigned a discrete uniform prior whose values are the M candidate neighbourhood matrices $(\boldsymbol{W}_G^1, \boldsymbol{W}_G^2, \ldots, \boldsymbol{W}_G^M)$ previously constructed. However, these candidate matrices do not have a natural ordering, thus when updating the choice of $\widetilde{\boldsymbol{W}}$ in the $(\mathrm{MC})^3$ algorithm, a new matrix is uniformly sampled from the s candidate matrices that are most similar to the current matrix. Here the similarity between two candidate neighbourhood matrices is measured by the percentage of edges that are removed from both matrices. s is a parameter controlling the acceptance rates and mixing of the update, and pilot runs suggest $s = 4$ is appropriate within this chapter.

6.4 Simulation study

This section quantifies the performance of the methodology outlined in Section 6.3 on simulated data under a range of scenarios, and compares its performance against two alternatives. The two existing models are those proposed by Lee et al. (2021) (denoted LM) and Rushworth et al. (2014) (denoted RL), where the former estimates the disease risk and local boundaries based on a single neighbourhood matrix that is estimated in advance, while the latter enforces a global level of spatial smoothness on the random effects surface, which does not involve any boundary identification. The aims of this study are to illustrate the improved risk estimation delivered by the proposed model in the presence of boundaries in the risk surface, and also to measure the accuracy of the identified boundaries.

6.4.1 Data generation

The study region is the set of $n = 257$ Intermediate Zones (IZs) that comprise the Greater Glasgow and Clyde Health Board region, which matches the motivating application described in Section 6.6. Simulated disease counts $\{Y_{it}\}$ are generated from the Poisson log-linear model (6.1) for $T = 7$ time periods, where the expected disease counts $\{E_{it}\}$ are based on the motivating study data, whose values range between 12.61 and 160.15 in a single IZ and year with a median of 74.09. In order to explore the impact of disease prevalence on estimation performance, these $\{E_{it}\}$ values are divided by the scale factors (SF) of 1, 2 and 4. Thus SF = 1 corresponds to the motivating data, SF = 2 corresponds to having a smaller number of expected counts, while SF = 4 represents a rare disease that has very small expected counts. Disease risks $\{R_{it}\}$ are generated by simulating spatio-temporal random effects $\{\phi_{it}\}$, and as previously described covariates are not included. The intercept term β_0 is fixed for all simulations at 0.01. The random effects are generated for each time period from a multivariate Gaussian distribution, whose precision matrix is defined by the Leroux prior (Leroux et al., 2000) with $\boldsymbol{Q}(\rho,\boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$. Here $\boldsymbol{W}$ is the border sharing neighbourhood matrix, the variance parameter τ^2 is fixed at 0.001, whereas the spatial dependence parameter ρ is varied in the simulation design to explore model performance with different levels of spatial correlation in the risk surface. Two values of $\rho = 0.9, 0.6$ are used to generate spatially correlated random effects $\boldsymbol{\phi}_t$, which respectively represent strong and moderate spatial dependence. Lower values of ρ are not considered because they rarely appear in real life applications.

To simulate spatial boundaries in the risk surface, a piecewise constant mean function is specified for the mean of $\boldsymbol{\phi}_t$, which we denote by $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{nt})$. The piecewise constant mean function contains three distinct values, so that each $\mu_{it} \in \{-Z, 0, Z\}$. Thus geographically neighbouring areal units that have the same mean value will have no boundary between them as their disease risks will be similar, while those pairs that have different mean values will have a boundary between their risks. The value of Z controls the size of the boundaries, and larger values represent boundaries that correspond to larger differences in disease risk between neighbouring areas. Three separate values $Z = 1, 0.5, 0.25$ are considered in the different scenarios of the simulation design, which respectively correspond to large, moderate and small boundaries in disease risk between neighbours. As touched on previously, in this study I only consider the case that the simulated boundaries remain constant during the study period, which is achieved by enforcing $\mu_{it} = \mu_{il}$ for all $t \neq l$. Figure 6.1 provides the template used for generating data with boundaries, which are shown by the blue dots. There are 338 boundaries in total, which correspond to approximately 50% of the set of edges in the study region. The template is designed based on the motivating respiratory disease data, where areas with low, medium and high SIR values are assigned a mean value of $-Z, 0, Z$ respectively, and are shaded in white, grey and black in the map. Table 6.1 summarises the 18 sub-scenarios considered in the simulation study, where the following quantities are varied: (i) varying boundary magnitudes via $Z = 1, 0.5, 0, 25$; (ii) varying disease prevalences via $\text{SF} = 1, 2, 4$; and (iii) varying degrees of spatial autocorrelation via $\rho = 0.9, 0.6$.

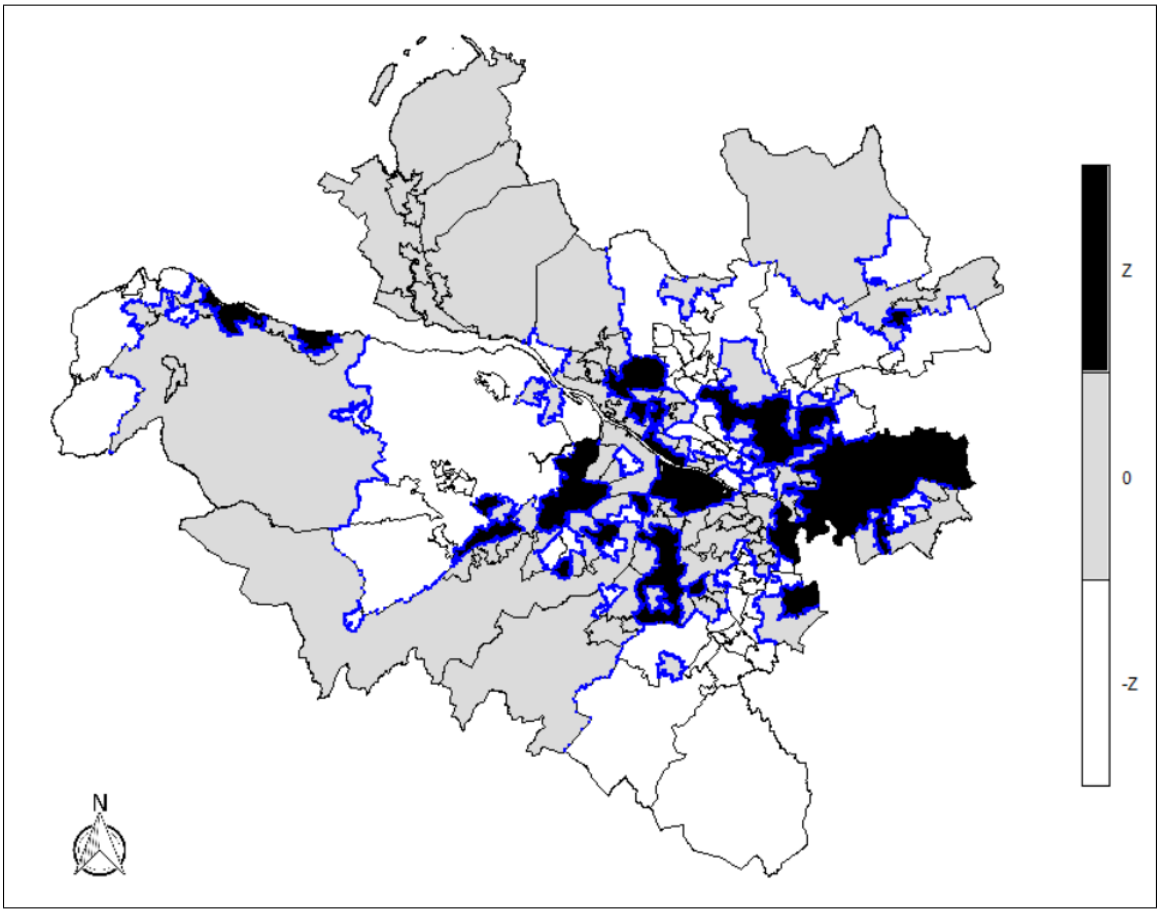


Figure 6.1: Locations of the true boundaries (which are highlighted by blue dots following the borders between the selected areal units) in the simulated random effects surfaces.

Table 6.1: Description of the scenarios considered in the simulation study.

Scenario	Z	SF	ρ
1	1	1	0.9
2	0.5	1	0.9
3	0.25	1	0.9
4	1	2	0.9
5	0.5	2	0.9
6	0.25	2	0.9
7	1	4	0.9
8	0.5	4	0.9
9	0.25	4	0.9
10	1	1	0.6
11	0.5	1	0.6
12	0.25	1	0.6
13	1	2	0.6
14	0.5	2	0.6
15	0.25	2	0.6
16	1	4	0.6
17	0.5	4	0.6
18	0.25	4	0.6

6.4.2 Results

One hundred simulated data sets are generated under each of the 18 scenarios shown in Table 6.1, and for each scenario the proposed model is compared to the LM model (Lee et al., 2021) and the RL model (Rushworth et al., 2014). These three models are fitted to each data set, and inference for each model is based on 2,000 samples obtained from generating 100,000 samples, the first 80,000 of which are discarded as the burn-in period and the remaining 20,000 are thinned by 10 due to limited computer memory capacity and to reduce autocorrelation. The convergence is assessed both by checking parameter trace plots and by Geweke (1992) diagnostics for a selection of the simulated data sets.

The results are summarised in Tables 6.2, 6.3 and 6.4. The accuracy of disease risk estimation is quantified by the root mean square error, RMSE $= \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{R}_{it} - R_{it})^2}$, and the coverage probabilities of the 95% credible intervals of the risk estimates. The overall fit of each model to each data set is summarised by the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), where a smaller value represents a better fitting model, and the effective number of independent parameters (p_d), where a smaller value indicates a more parsimonious model. The correctness of the boundary identification is measured by the receiver-operating characteristic (ROC) curves and the area under the ROC curve, abbreviated AUC (Bradley, 1997, Hanley and McNeil, 1982), which are based on the sensitivity and specificity of the proposed model at identifying true boundaries/non-boundaries. Firstly, to estimate this I compute the probability of each edge being removed (i.e. the probability of each element $\tilde{w}_{ij}$ in $\widetilde{\boldsymbol{W}}$ being 0 for all $i \sim j$) across all the posterior samples of $\widetilde{\boldsymbol{W}}$, that is

$$p(\tilde{w}_{ij} = 0|\boldsymbol{Y}) = \frac{1}{G}\sum_{g=1}^{G}(1 - \tilde{w}_{ij}^{(g)}), \text{ for all } i \text{ geographically adjacent to } j, \qquad (6.9)$$

where $\tilde{w}_{ij}^{(g)}$ represents the value of element $\tilde{w}_{ij}$ in the g_{th} posterior sample of $\widetilde{\boldsymbol{W}}$. Then I define a threshold p^* for identifying a boundary, and compare it to $p(\tilde{w}_{ij} = 0|\boldsymbol{Y})$. If $p(\tilde{w}_{ij} = 0|\boldsymbol{Y}) \geq p^*$ then a boundary is identified between the random effects of the two adjacent areas (i, j), whereas if $p(\tilde{w}_{ij} = 0|\boldsymbol{Y}) < p^*$ no boundary is detected between them. The sensitivity and specificity of the boundary identification are computed at different values of p^*, which is varied from 0 to 1.01 at intervals of 0.01 in the study, and a ROC curve is a plot of sensitivity against specificity. The sensitivity is computed as the percentage of the true boundaries that are correctly identified, while the specificity is the percentage of the non-boundaries correctly identified. The closer an ROC curve is to the upper left corner, the more accurate is the identification. Thus in order to summarise the "overall" location of the entire ROC curve, the area under the curve (AUC) is computed, which provides an aggregate measure of boundary identification performance across all possible threshold values. It takes values from 0 to 1, where an AUC of 1 corresponds to perfectly accurate boundary identification, an AUC of 0.5 implies a random identification such that the ROC curve will fall on the diagonal (45-degree line), and an AUC of 0 indicates perfectly inaccurate identification. Note that the ROC curve and AUC statistics are only available for the proposed model, because the neighbourhood matrix is fixed when estimating disease risk in the other two models LM and RL.

Table 6.2 displays the RMSE, 95% coverage probabilities, DIC and p_d associated with each model and each scenario when the spatial dependence in the simulated data is strong ($\rho = 0.9$). The table clearly illustrates that estimating an appropriate neighbourhood matrix that accounts for the boundaries in the data provides improved estimation compared to simply using the border sharing $\boldsymbol{W}$. The RL model overall performs the worst of three in terms of the largest RMSE, DIC and p_d, which is due to it enforcing global spatial smoothing of disease risk across the region, so it does not allow neighbouring areas to have very different values (or boundaries), resulting in poorer risk estimation and model fit. Both the proposed model and the LM model perform very well in terms of risk estimation, but the former seems to have slightly smaller RMSE values and higher coverages. However, the proposed model exhibits a higher DIC and p_d than the LM model in most scenarios. This is likely to be because the neighbourhood matrix in the LM model is estimated in advance via an optimisation algorithm, so it is fixed during the estimation procedure. In contrast, our model estimates $\widetilde{\boldsymbol{W}}$ from a set of possible candidate values as part of the modelling approach, which thus results in an increase in p_d and DIC. Credible interval coverages are generally good for all three models, with values varying between 0.92 and 0.96. The improved performance of our model becomes less noticeable when the magnitude of the boundaries decreases (Z gets smaller) compared to the RL model, which is likely the result of less accurate boundary identification as shown in Table 6.4. For all models, RMSE increases as the disease prevalence decreases (SF > 1), which is due to a reduction in the amount of data. Table 6.3 summarises model performance when reducing the spatial dependence ($\rho = 0.6$) in the simulated data. The results from all metrics for each model and each scenario are very close to those in Table 6.2, thus the main findings outlined above are unaffected by reducing the level of spatial dependence from $\rho = 0.9$ to $\rho = 0.6$. The proposed model and the LM model produce a better estimation of disease risk and model fit than the globally smooth RL model across all scenarios.

An ROC curve is a two-dimensional depiction of the boundary identification performance. Thus for ease of presentation, only the AUC statistics are presented here rather than the full ROC curves. Table 6.4 displays the median AUC values as well as the corresponding 95% credible intervals across the set of ROC curves calculated for each scenario for the proposed model. In general, the model performs well in identifying the true boundaries and non-boundaries. When the size of the boundaries is large ($Z = 1$), the model exhibits outstanding boundary identification, with median AUC values close to the maximum value of 1. When the boundaries are less pronounced ($Z < 1$) and the disease prevalence decreases ($\text{SF} > 1$), the model performs slightly less well but the AUC values are still relatively high, which range between 0.818 and 0.951. The exception to this is the scenario when the disease is rare ($\text{SF} = 4$) and the boundaries are very small in magnitude ($Z = 0.25$), which obtains lower AUC values (around 0.75). The reason for this is that in this scenario the boundaries are difficult to correctly identify based on their small size and small numbers of disease cases. Finally, the AUC statistics obtained when $\rho = 0.9$ are very similar to those obtained when $\rho = 0.6$, which also suggest that the ability of the model to identify the correct boundaries is robust to the level of spatial dependence in the data.

Table 6.2: Median values of the RMSE, 95% credible interval coverages associated with the estimated risks, Deviance Information Criterion (DIC), and the effective number of parameters (p_d) for each model and scenario when $\rho = 0.9$ is used to simulate the spatial random effects for each time period. Here LM and RL refer to the models proposed by Lee et al. (2021) and Rushworth et al. (2014) respectively.

Metric	Z	SF	Model		
			Proposed	LM	RL
RMSE	1	1	0.100	0.101	0.121
	0.5	1	0.074	0.075	0.089
	0.25	1	0.062	0.062	0.068
	1	2	0.131	0.132	0.161
	0.5	2	0.098	0.099	0.113
	0.25	2	0.080	0.081	0.084
	1	4	0.174	0.174	0.209
	0.5	4	0.127	0.128	0.141
	0.25	4	0.103	0.103	0.104
Coverage probability	1	1	0.963	0.960	0.954
	0.5	1	0.958	0.956	0.954
	0.25	1	0.951	0.944	0.948
	1	2	0.961	0.957	0.952
	0.5	2	0.952	0.949	0.951
	0.25	2	0.949	0.931	0.942
	1	4	0.953	0.951	0.949
	0.5	4	0.950	0.941	0.947
	0.25	4	0.953	0.919	0.934
DIC	1	1	13714.99	13701.59	14123.56
	0.5	1	13473.27	13456.07	13770.36
	0.25	1	13268.14	13259.73	13454.56
	1	2	12328.81	12319.42	12692.48
	0.5	2	12099.90	12087.15	12342.82
	0.25	2	11908.75	11896.97	12060.67
	1	4	10948.64	10940.93	11254.94
	0.5	4	10736.03	10723.66	10924.71
	0.25	4	10567.33	10542.44	10675.29
p_d	1	1	992.45	988.89	1345.70
	0.5	1	721.27	730.13	1007.26
	0.25	1	521.50	509.29	653.65
	1	2	846.93	844.98	1175.94
	0.5	2	599.94	605.26	819.69
	0.25	2	433.99	398.49	499.04
	1	4	712.76	720.97	991.77
	0.5	4	507.93	495.56	643.36
	0.25	4	376.80	313.91	366.86

Table 6.3: Median values of the RMSE, 95% credible interval coverages associated with the estimated risks, Deviance Information Criterion (DIC), and the effective number of parameters (p_d) for each model and scenario when $\rho = 0.6$ is used to simulate the spatial random effects for each time period. Here LM and RL refer to the models proposed by Lee et al. (2021) and Rushworth et al. (2014).

Metric	Z	SF	Model		
			Proposed	LM	RL
RMSE	1	1	0.101	0.101	0.122
	0.5	1	0.075	0.076	0.089
	0.25	1	0.062	0.063	0.069
	1	2	0.132	0.133	0.160
	0.5	2	0.098	0.099	0.113
	0.25	2	0.080	0.081	0.084
	1	4	0.174	0.175	0.210
	0.5	4	0.128	0.129	0.142
	0.25	4	0.103	0.103	0.104
Coverage probability	1	1	0.963	0.961	0.954
	0.5	1	0.957	0.956	0.954
	0.25	1	0.946	0.939	0.947
	1	2	0.959	0.957	0.953
	0.5	2	0.951	0.949	0.952
	0.25	2	0.948	0.931	0.943
	1	4	0.953	0.952	0.950
	0.5	4	0.951	0.940	0.947
	0.25	4	0.949	0.917	0.932
DIC	1	1	13717.41	13710.03	14122.91
	0.5	1	13461.56	13450.71	13763.70
	0.25	1	13259.48	13250.67	13437.23
	1	2	12319.30	12311.37	12686.15
	0.5	2	12090.22	12083.78	12334.35
	0.25	2	11921.21	11906.64	12068.50
	1	4	10932.32	10923.13	11236.89
	0.5	4	10730.95	10715.64	10914.66
	0.25	4	10576.64	10542.19	10681.83
p_d	1	1	992.40	987.48	1346.29
	0.5	1	717.32	727.01	1005.38
	0.25	1	517.14	503.34	654.19
	1	2	845.26	847.54	1174.93
	0.5	2	609.95	606.38	819.71
	0.25	2	437.51	402.88	502.37
	1	4	721.48	720.27	991.15
	0.5	4	514.07	496.90	644.61
	0.25	4	372.89	312.70	365.94

Table 6.4: Median values of the area under the ROC curve (AUC) for boundary identification for the proposed model for each scenario. Values in brackets correspond to the 95% credible intervals.

Z	SF	ρ	AUC
1	1	0.9	0.977 (0.960, 0.985)
0.5	1	0.9	0.951 (0.933, 0.965)
0.25	1	0.9	0.876 (0.846, 0.902)
1	2	0.9	0.966 (0.944, 0.978)
0.5	2	0.9	0.919 (0.896, 0.942)
0.25	2	0.9	0.819 (0.790, 0.856)
1	4	0.9	0.945 (0.926, 0.962)
0.5	4	0.9	0.877 (0.850, 0.903)
0.25	4	0.9	0.748 (0.706, 0.778)
1	1	0.6	0.977 (0.960, 0.986)
0.5	1	0.6	0.949 (0.925, 0.965)
0.25	1	0.6	0.869 (0.841, 0.900)
1	2	0.6	0.966 (0.953, 0.978)
0.5	2	0.6	0.918 (0.894, 0.942)
0.25	2	0.6	0.818 (0.786, 0.859)
1	4	0.6	0.944 (0.920, 0.959)
0.5	4	0.6	0.875 (0.836, 0.900)
0.25	4	0.6	0.748 (0.704, 0.783)

6.5 Sensitivity analysis

The model developed here uses an Inverse-Gamma(1,0.01) prior for the variance parameter τ^2. To assess the impact of the prior for τ^2 on model performance, I compare the choice with two alternatives, which are Inverse-Gamma(0.001,0.001) and Inverse-Gamma(0.5,0.0005). One hundred simulated data sets are generated as described in Section 6.4 for each value of $Z = 1, 0.5, 0.25$, where $\rho = 0.9$ is used to simulate the random effects $\boldsymbol{\phi}_t$ at each time period and the expected numbers of cases are taken from the motivating data. The proposed model is fitted to each data set using the three different choices of Inverse-Gamma (IG) prior distribution for τ^2, and the results are summarised in Figures 6.2, 6.3 and 6.4, which display boxplots of RMSE, 95% coverage probabilities for risk estimates, DIC, p_d and the AUC over all simulated data sets.

The figures show that the model yields very similar results in terms of both risk estimation and boundary identification using the three different priors. A slight difference is observed in the scenario of $Z = 0.25$, where IG(1,0.01) yields slightly higher p_d values with a median of 538 compared to 530 for the other two priors. Therefore the proposed model appears to be robust to the choice of the hyperparameters of the prior Inverse-Gamma distribution for τ^2.

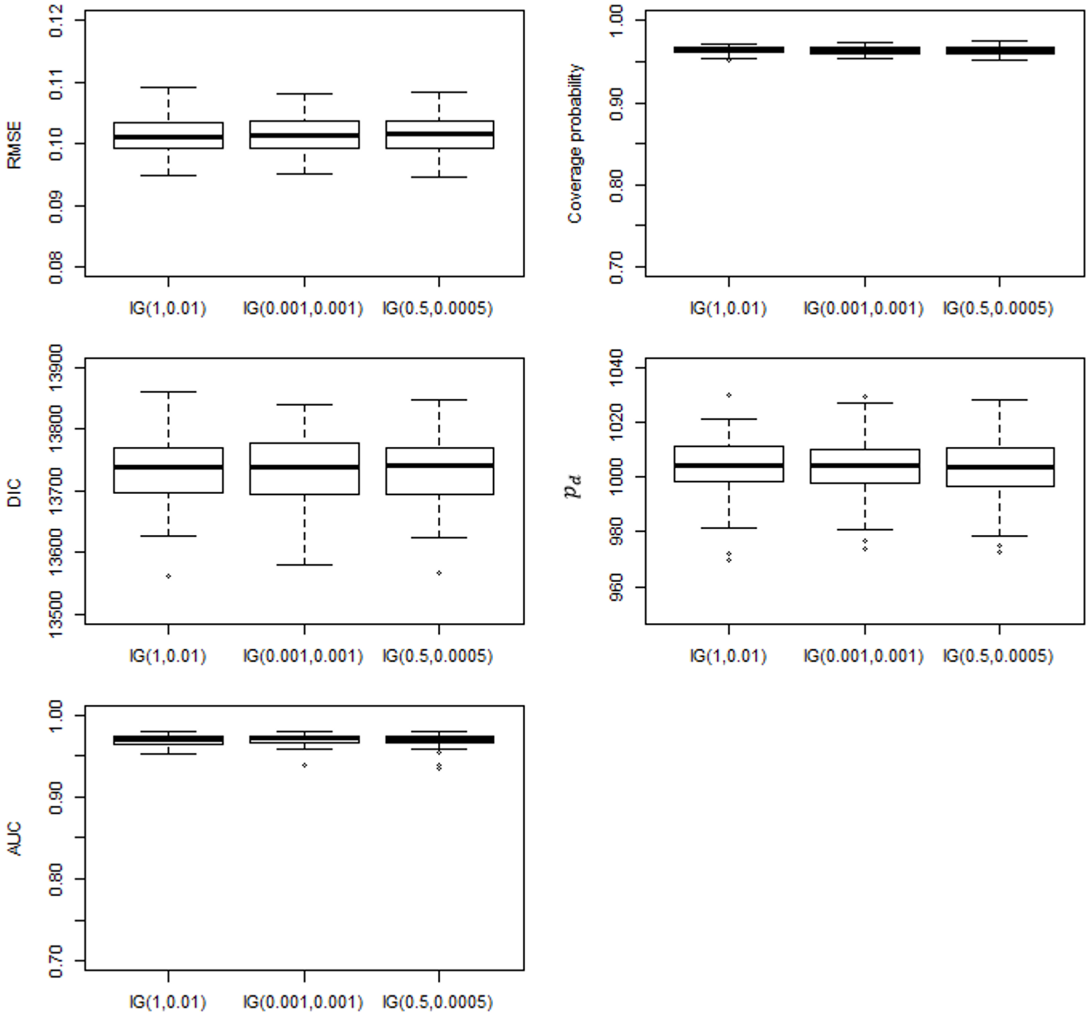


Figure 6.2: Summary of the simulation results from changing the hyperparameters of the Inverse-Gamma (IG) prior distribution for τ^2 when $Z = 1$.

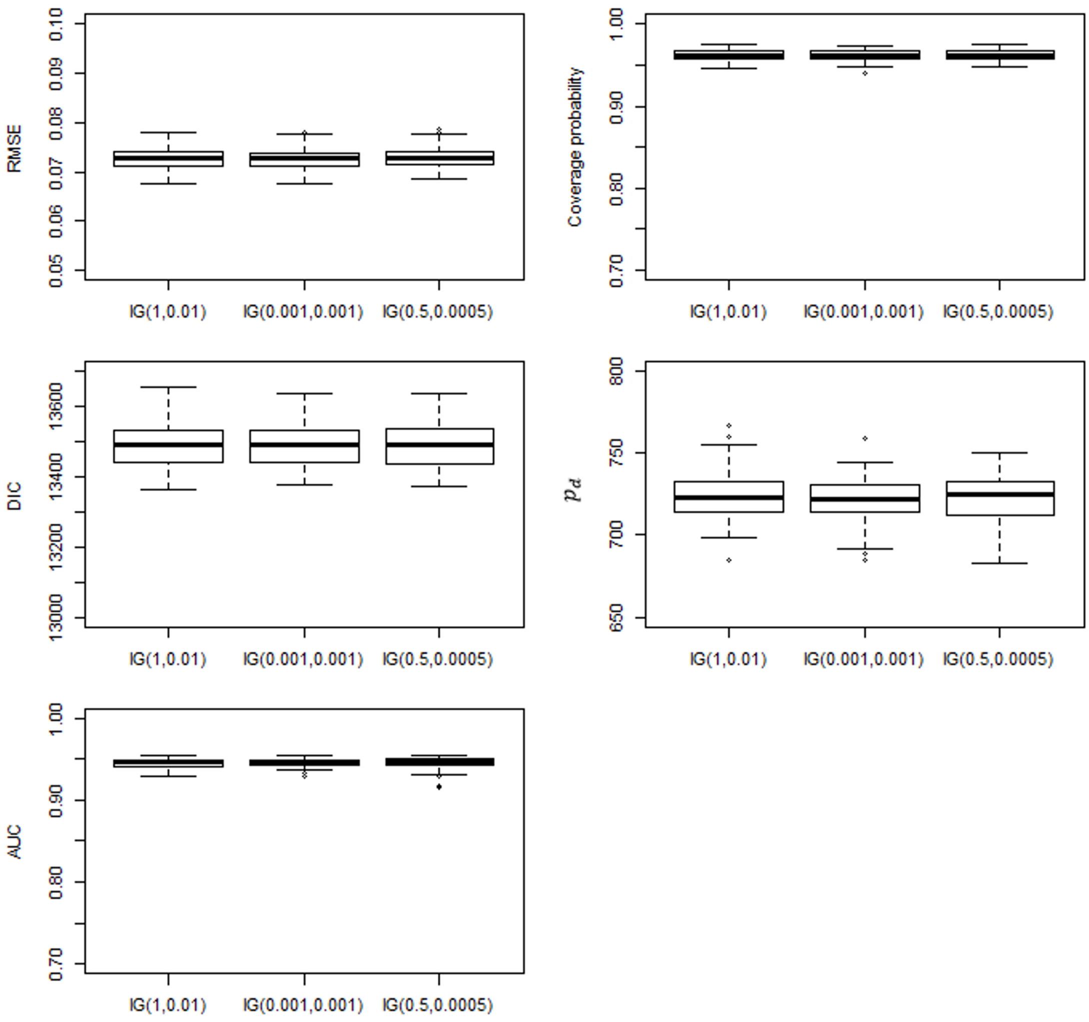


Figure 6.3: Summary of the simulation results from changing the hyperparameters of the Inverse-Gamma (IG) prior distribution for τ^2 when $Z = 0.5$.

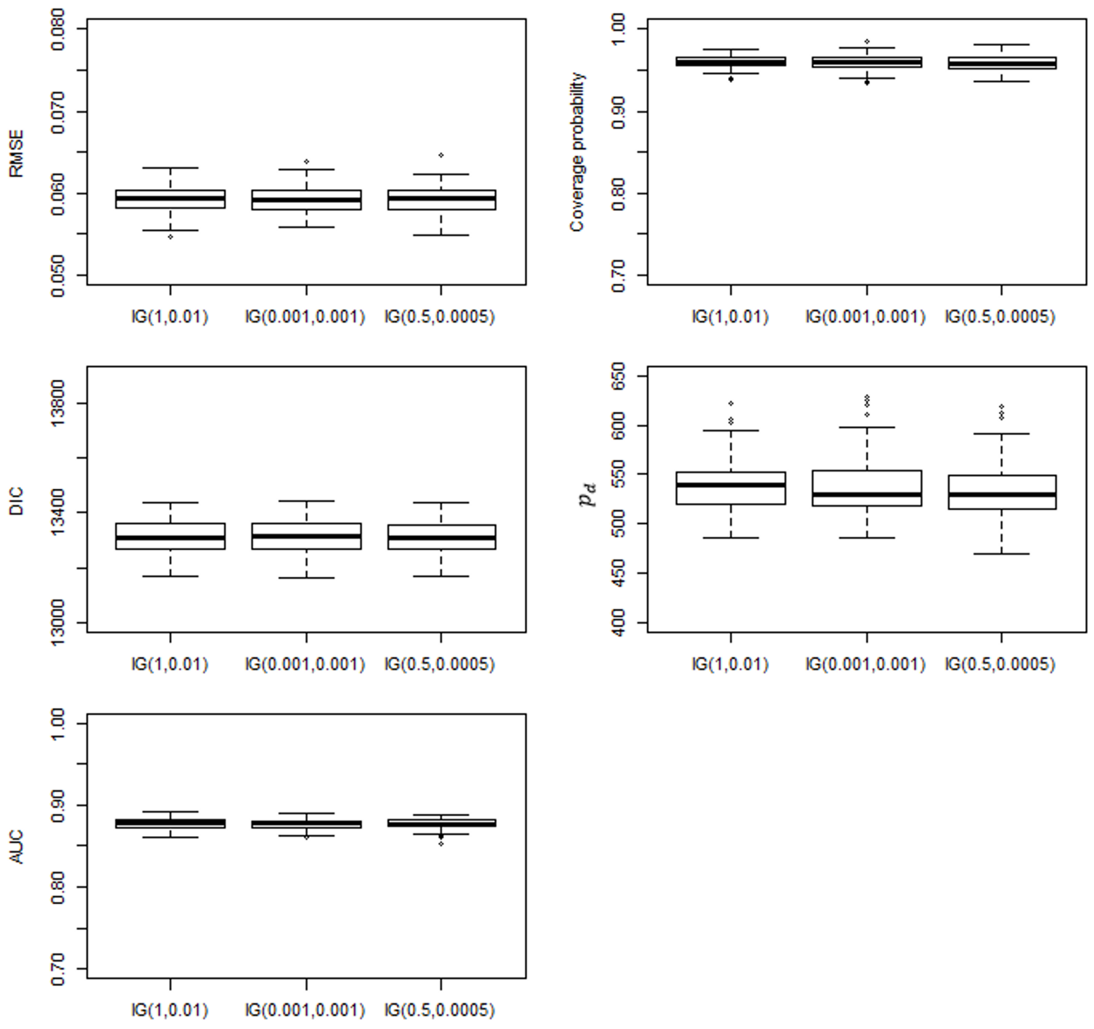


Figure 6.4: Summary of the simulation results from changing the hyperparameters of the Inverse-Gamma (IG) prior distribution for τ^2 when $Z = 0.25$.

6.6 Glasgow respiratory disease study results

This section continues to analyse the respiratory admission data presented in Section 5.2.1. As in the previous chapters, the study region is the Greater Glasgow and Clyde Health Board displayed in Figure 3.1, which consists of $n = 257$ non-overlapping Intermediate Zones (IZs). The data in each IZ are collected for 7 years from 2011 to 2017. The disease data $\boldsymbol{Y} = \{Y_{it}\}$ are the yearly counts of hospital admissions with a primary diagnosis of respiratory disease for $i = 1, \dots, n(n = 257)$ IZ in year $t = 1, \dots, T(T = 7)$, and range between 17 and 282. The expected numbers of respiratory hospitalisations $\boldsymbol{E} = \{E_{it}\}$ are calculated for each year and IZ to adjust for different population sizes and demographic structures across the IZs using indirect standardisation. The standardised incidence ratio (SIR) is the simplest measure of disease risk calculated by $\text{SIR}_{it} = \frac{Y_{it}}{E_{it}}$. The middle panel in Figure 5.1 displays the spatial pattern in the SIR for 2017, and shows that the respiratory disease risks are highest in the

East End of Glasgow (the east of the map) and along the southern bank of the River Clyde, which contains a number of heavily deprived areas. It also shows that the spatial risk surface is not globally smooth and contains numerous pairs of geographically adjacent areas where there appear to be discontinuities in their risks, suggesting the presence of boundaries.

6.6.1 Model choice and inference

The methodology outlined in Section 6.3 is applied to the respiratory disease data in the Greater Glasgow and Clyde Health Board during the time period from 2011 to 2017. For comparison purpose, the model of Lee et al. (2021) (denoted LM) and the global smoothing model of Rushworth et al. (2014) (denoted RL) are also fitted to the data to observe which one best fits the data and hence will likely produce the best estimates of disease risk. The goals in analysing these data are two-fold: on the one hand, providing the best estimate of the spatio-temporal patterns in respiratory disease risk, on the other hand, estimating the locations of any boundaries in the spatial risk surface, so that the areas that exhibit excessively high risks compared to their neighbours can be identified. Posterior inference for all models is based on five independent Markov chains, where each chain is run for 100,000 samples with a burn-in period of 80,000 and the remaining 20,000 samples are then thinned by 5, yielding a total of 20,000 samples across all five chains.

6.6.2 Overall model fit

Table 6.5 summarises the overall fit of each model to the data by presenting both the DIC and the effective number of independent parameters p_d. It shows that the proposed model fits the data better than the global smoothing RL model in terms of DIC, with a value of 14,080 compared to 14,141. Our model also has a markedly smaller p_d than the RL model, which is due to its lower estimate of the random effects variance τ^2 as seen in Table 6.5. Our model does not allow any smoothing of random effects between pairs of geographically adjacent IZs that exhibit largely different risks, which thus makes the random effects smooth more strongly elsewhere in the spatial surface and reduces the amount of variation between $\{\phi_{it}\}$. This suggests a greater level of penalisation of the random effects and hence leads to a reduction in the overall p_d. The proposed model has a slightly higher DIC and p_d value than the LM model, which is due to it estimating $\widetilde{\boldsymbol{W}}$ in the modelling process rather than fixing it when estimating disease risk as in the LM model.

Table 6.5: A summary of the overall fit of each model and the estimated (posterior median) random effects variance τ^2.

	Proposed	LM	RL
DIC	14,080.41	14,072.72	14,140.90
p_d	849.69	830.17	1054.90
τ^2	0.015	0.014	0.070

6.6.3 Temporal trends in disease risk

The proposed model and the LM model have the most similar risk estimates, with a mean absolute difference in the posterior median risk estimates of 0.010 compared to 0.030 (Proposed vs RL) and 0.034 (RL vs LM) over all years and IZs. Here I present the estimated risks from the proposed model in Figure 6.5, because it performs a better fit to the data than the RL model, and it also has the advantage of being able to quantify the uncertainty in the locations of boundaries in the spatial risk surface through the posterior distribution of $\widetilde{\boldsymbol{W}}$ compared to the LM model. Figure 6.5 displays boxplots of the risk estimates from all the areal units over time. These risk estimates are broadly similar to those obtained in Section 5.6. An increasing trend in risk is observed for the entire time period. In 2011 the average risk across Greater Glasgow is 1.10, suggesting that on average respiratory disease risk in Greater Glasgow is about 10% higher than the Scottish average. This rises to 1.28 in the final year 2017, which is thus 28% higher than the overall Scotland average. In addition, the health inequalities across Glasgow in respiratory disease have widened over time, as the difference between the largest and smallest disease risk is 1.36 in 2011 and 2.13 in 2017.

There are some small evolution in the estimated spatial risk patterns over the 7-year period, with the Pearson's correlation coefficient between any pair of years ranging between 0.90 and 0.98. Figure 6.6 presents the spatial patterns of the estimated disease risks across Greater Glasgow in 2011, 2014 and 2017. The increasing trend in risk over time can also be seen in the figure as the shading gets darker from 2011 to 2017. In all three maps there are common IZs having darker shading, indicating higher respiratory disease risk. The areas with higher risks tend to be in the east (e.g. Easterhouse, Parkhead), and the north of Glasgow city centre (e.g. Springburn, Possilpark), as well as in the south of the Clyde river (e.g. Nitshill, Priesthill, Govan), the north-west (e.g. Drumchapel) and the south-west (e.g. Castlemilk). In contrast, the areas of lower risks are located in the West End (just the north of the Clyde river) and the south-west of the city centre, such as Hyndland, Kelvinside, Jordanhill and Newton Mearns, and also in the far south such as Eaglesham. These results show that people living in deprived areas are more likely to be hospitalised for respiratory disease than those living in affluent areas, which are consistent with the findings reported in the previous chapters. Furthermore, the maps also suggest that the risks in the areas that have high risks at the beginning of the time period appear to increase more quickly over time than in the areas with low risks at the start.

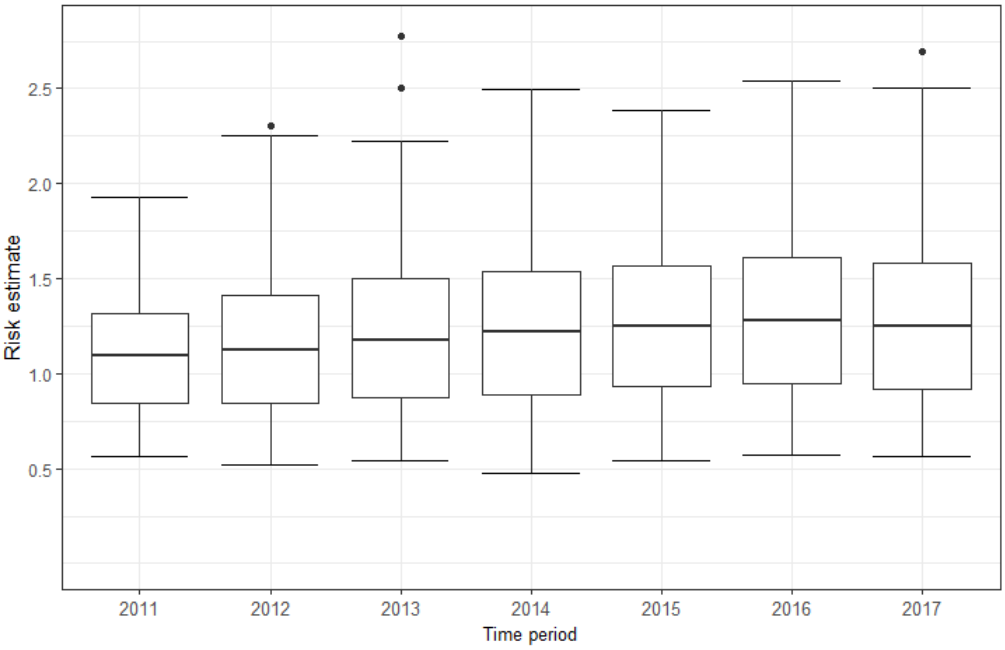
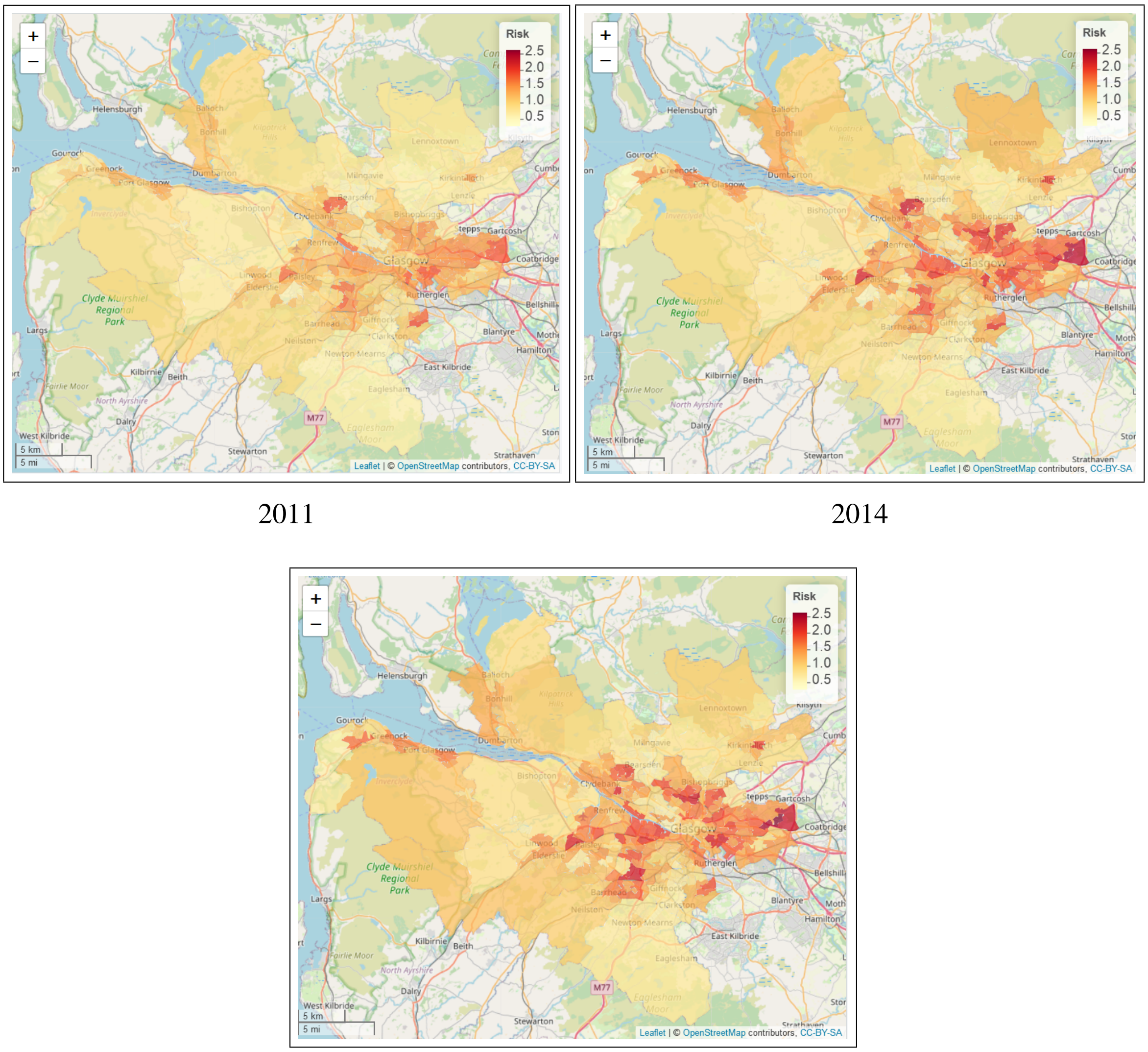


Figure 6.5: Boxplots of the risk estimates (posterior median) for all the areal units over time from the proposed model.



2011

2014

2017

Figure 6.6: Maps of the respiratory disease risk estimates (posterior median) in Greater Glasgow for 2011, 2014 and 2017 from the proposed model.

6.6.4 Boundary identification

The proposed model provides additional inference on the locations of boundaries in the risk surface, which separate areas that are geographically adjacent and exhibit very different risks. Identifying these boundaries is important for social epidemiologists because their locations are likely to reflect "*underlying biological, physical, and/or social processes*" (Jacquez et al., 2000). I quantify the evidence for a boundary between adjacent areas (i, j) using $p(\tilde{w}_{ij} = 0|\boldsymbol{Y})$, that is the probability of element $\tilde{w}_{ij}$ in $\widetilde{\boldsymbol{W}}$ being 0, which is computed by equation (6.9). Table 6.6 summarises the number of spatial boundaries identified by the model, based on $p(\tilde{w}_{ij} = 0|\boldsymbol{Y})$ values greater than or equal to a threshold p^*. Three different threshold values $p^* = 1, 0.99, 0.975$ are considered here, where $p^* = 0.99$ was used by Lu and Carlin (2005) and Rushworth et al. (2017). The results indicate that the boundaries are almost unchanged by reducing the value of p^* from 1 to 0.975. Values of $p^* = 1, 0.99$ identify the same number of boundaries (41% of all the edges in this study region), with complete agreement between their locations, and $p^* = 0.975$ identifies one more boundary than the other two threshold values. Figure 6.7 maps the average risk estimates (posterior median) over 2011-2017 from the proposed model. The grey-scale lines denote the locations of the boundaries that are identified using a threshold of $p(\tilde{w}_{ij} = 0|\boldsymbol{Y}) = 1$ in equation (6.9). These boundaries are detected in the random effects surface, but since covariates are not included in the model, the random effects and risk surfaces have the same spatial structure as $R_{it} = \exp(\beta_0 + \phi_{it})$, and thus any boundaries identified also relate to disease risk. The identified boundaries are shaded based on their size of the differences in disease risk between geographically adjacent IZs. Specifically, I compute the time averaged absolute differences in disease risk between adjacent IZs for each identified boundary. Then I rank the boundaries from smallest to largest by their mean absolute differences, and normalize the ranks by dividing them by the total number of boundaries. This ensures the ranks after transformation fall in the range $[0, 1]$ and the transformed values suggest the relative strength of the identified boundaries. Therefore, boundaries with darker shading in Figure 6.7 correspond to larger differences in disease risk between pairs of neighbouring IZs, which may need more attention and investigation from health authorities. According to the model design, the identified boundaries are common to all time periods.

Figure 6.7 suggests that the respiratory disease risk surface in Greater Glasgow is far from being globally spatially smooth, which is the reason why the proposed model performs better than the global smoothing model. Boundaries that have been identified appear to mainly correspond to sizeable changes in disease risk between adjacent IZs. The model identifies the largest boundaries in the north-west between Drumchapel and its neighbour Bearsden, in the north between Springburn and Bishopbriggs as well as in the East End of the city between Easterhouse and Garrowhill. The mean risk in the deprived Drumchapel (which is 2.05) is about three times as high as that in the affluent Bearsden (which is 0.72). Springburn and Easterhouse have vastly higher mean risks of 1.92 and 2.17 compared to 0.8 for Bishopbriggs and 1.07 for Garrowhill. Other prominent boundaries are found in the south-west of the city and also in the south-east. For example, Househillwood is separated from its neighbour Roughmussel with a mean risk of 2 compared to 0.98, and the mean risk in Castlemilk is about 60% higher than its surrounding area Carmunnock. I also notice that some boundaries are near motorways, railways or big roads (e.g. the North Clyde Line, A82). This is probably because such physical barriers prevent people who live on either side from mixing and thus may result in different population behaviour. Note that adjacencies are not assumed between pairs of areas across the river in the border sharing $\boldsymbol{W}$, therefore boundaries cannot be identified between these areas.

Table 6.6: A summary of the number of boundaries/non-boundaries identified at different values of the threshold p^*.

Value of p^*	Number of boundaries	Number of non-boundaries
1	277	394
0.99	277	394
0.975	278	393

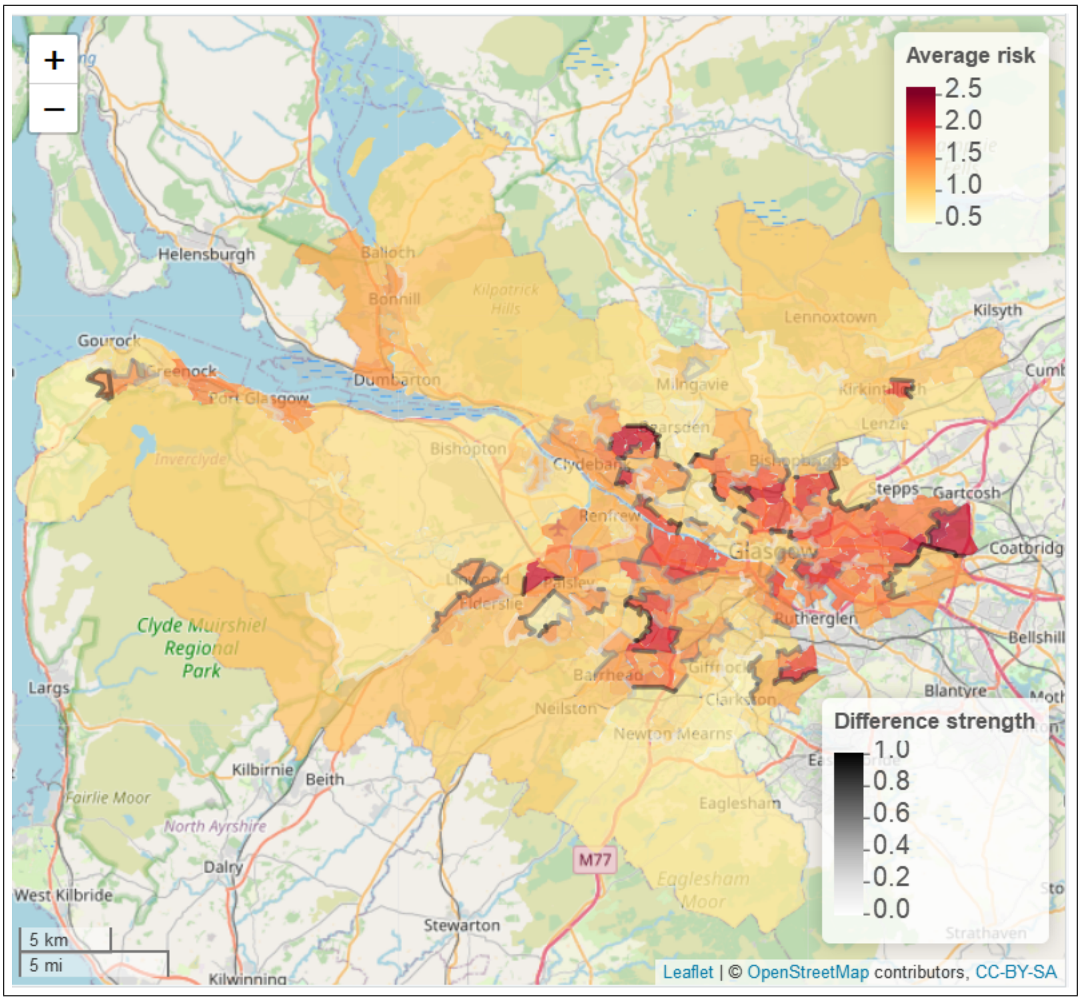


Figure 6.7: A map of the average respiratory disease risk surface in Greater Glasgow over 2011-2017 from the proposed model. The grey-scale lines in the map correspond to the boundaries that have been identified using a threshold of $p^* = 1$ in equation (6.9). Boundaries with darker shading represent larger differences in risk between geographically adjacent IZs.

6.6.5 Sensitivity analysis

In order to examine the sensitivity of the above results to the choice of hyperparameters, the Inverse-Gamma$(1, 0.01)$ prior for the random effects variance parameter τ^2 is changed to Inverse-Gamma$(0.001, 0.001)$ and Inverse-Gamma$(0.5, 0.0005)$. Figure 6.8 presents scatter plots of the risk estimates over 2011-2017 among different choices of prior Inverse-Gamma distribution for τ^2. It shows that the estimated risks are very similar for all three prior choices, with data points lying on the diagonal line and the correlation coefficients between the risk estimates for any two priors being close to 1. Therefore, the results reported above are robust to the choice of the hyperpriors for τ^2.

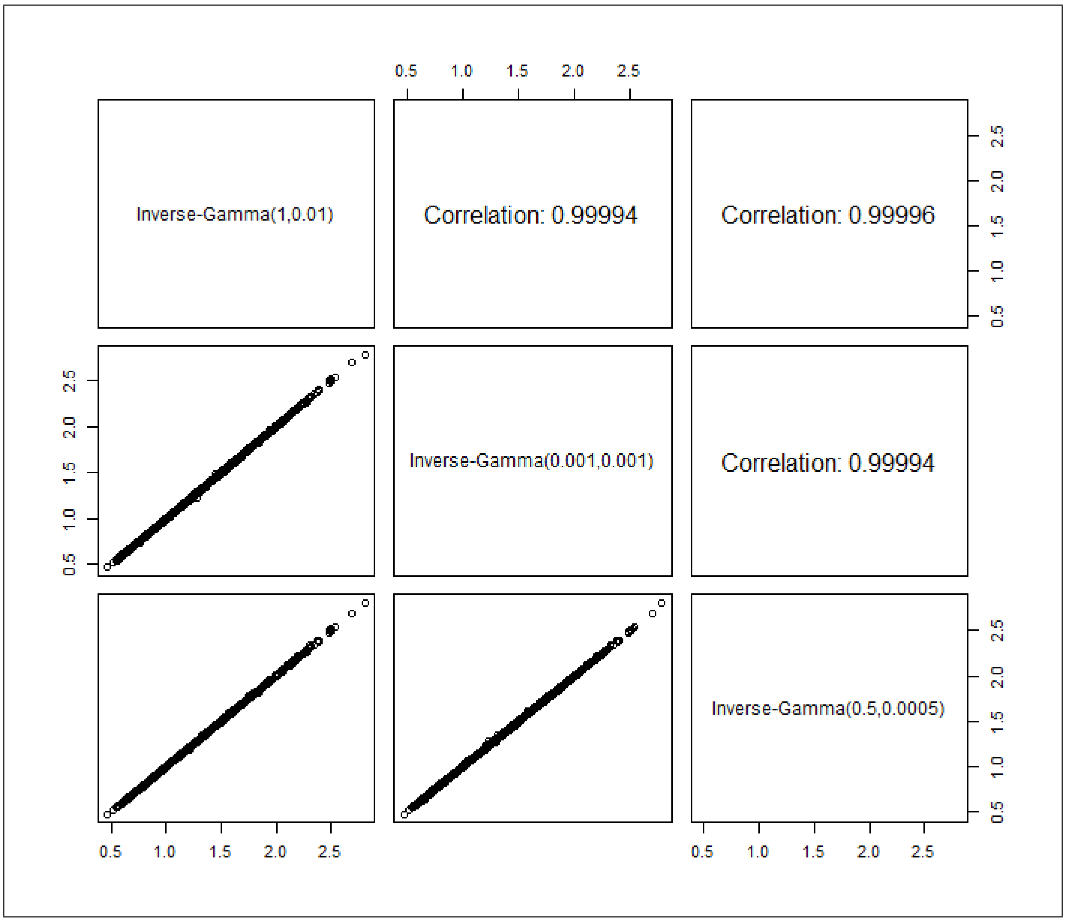
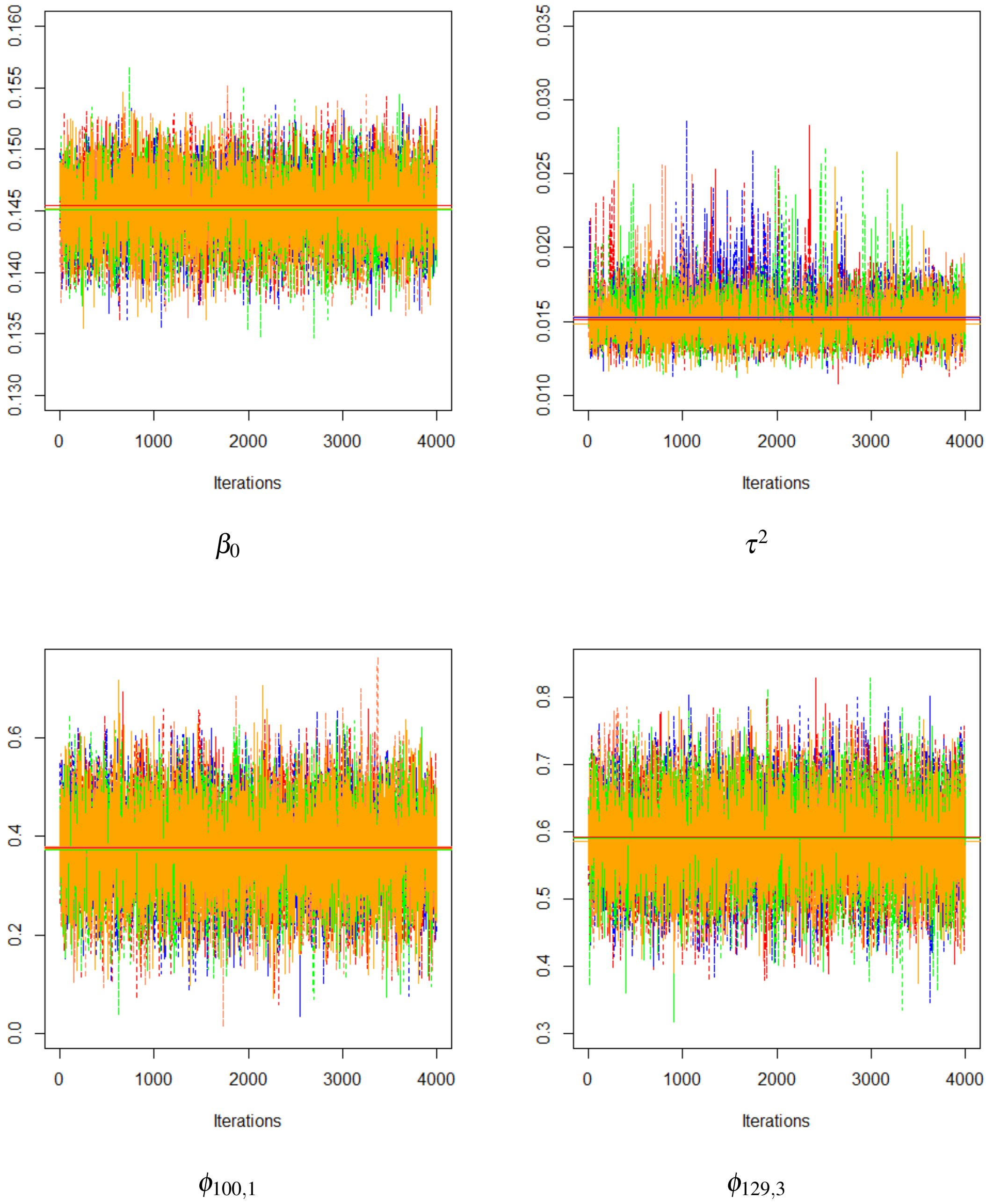


Figure 6.8: Scatter plots of the estimated risks (posterior median) between different choices of prior Inverse-Gamma distribution for τ^2.

6.6.6 Convergence diagnostic

The convergence of the posterior distributions is diagnosed both by examining parameter trace plots and by Gelman–Rubin diagnostic (Gelman et al., 1992). It is infeasible to check the convergence for all parameters in practice because there are a large number of parameters in the model. Therefore here I only select four parameters, which are $(\beta_0, \tau^2, \phi_{100,1}, \phi_{129,3})$, to check their convergence. Figure 6.9 shows trace plots of the posterior samples for $(\beta_0, \tau^2, \phi_{100,1}, \phi_{129,3})$, where each Markov chain is represented in a different color. The figure shows that there is no clear pattern in the trace plots for all selected parameters, which suggest that all the chains appear to have converged. In addition, the Gelman-Rubin diagnostic is also used to check the convergence for multiple chains, with a value less than 1.1 indicating convergence of the chains. Here the Gelman-Rubin statistics for the selected parameters are all smaller than 1.1 with a maximum value of 1.01, which mean that the posterior samples appear to have converged.



β_0 τ^2

$\phi_{100,1}$ $\phi_{129,3}$

Figure 6.9: Trace plots of the posterior samples for selected parameters $(\beta_0, \tau^2, \phi_{100,1}, \phi_{129,3})$ from the proposed model.

6.7 Discussion

In this chapter I propose an approach for estimating the disease risk pattern over time by using CAR priors, and also identifying the locations of boundaries in the risk surface by estimating the neighbourhood matrix for the data rather than assuming it is fixed on the basis of geographical adjacency. The model is an extension of Lee et al. (2021) by allowing to account for the uncertainty in the neighbourhood matrix in the modelling process. The methodology first uses the graph-based optimisation algorithm developed by Lee et al. (2021) to obtain multiple possible candidate neighbourhood matrices, which represent a range of possible boundary structures in the data. Then a Bayesian spatio-temporal model is fitted to the data, in which the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ is treated as a random quantity to be estimated from the set of candidates previously constructed. To perform inference, a Metropolis-coupled Markov chain Monte Carlo algorithm is used to yield posterior samples for model parameters, which overcomes the multimodality issue in the posterior distribution.

The simulation study in Section 6.4 has shown strong evidence that the model developed here outperforms the global smoothing model developed by Rushworth et al. (2014) in the presence of boundaries, in terms of both risk estimation and model fit. This improved performance results from our model having the flexibility to represent a range of localised spatial autocorrelation structures that account for the risk boundaries. In contrast, the global smoothing model induces spatial autocorrelation structure based on geographical adjacency and hence smooths the disease risk over such boundaries, which leads to poorer estimation of disease risk. The study also shows that the proposed model and the LM model developed by Lee et al. (2021) produce very similar results in terms of risk estimation. This is probably because the candidate neighbourhood matrices $\left(\boldsymbol{W}_G^1, \boldsymbol{W}_G^2, \ldots, \boldsymbol{W}_G^M\right)$ obtained in stage one do not differ vastly from each other. Although the proposed model has the advantage of allowing for uncertainty in $\widetilde{\boldsymbol{W}}$ in theory, it does not make a large difference in risk estimation compared to the LM model. The model developed here is also successful at identifying the locations of true boundaries in simulated data, with AUC statistics close to 1 for a set of different scenarios. The model tends to obtain higher AUC values if the magnitude of the boundaries gets larger and the disease in question is not rare.

The motivating application also establishes the superiority of the proposed model compared to the global smoothing model. Our model produces a better model fit with a smaller number of effective parameters, due to its increased levels of smoothing in locations where boundaries are not present. Although the model has a marginally higher DIC value than the LM model, it is able to measure the uncertainty of each edge being identified as a boundary. Figure 6.7 provides substantial evidence of boundaries in the unexplained spatio-temporal risk pattern for respiratory disease in Greater Glasgow. Most of the identified boundaries correspond to sizeable discontinuities in the risk surface, and are very likely to occur where poor and wealthy areas border or be close to geographical barriers such as railways or big roads. The application results appeared to be robust when the proposed model was re-run for another $M = 100$ different candidate neighborhood matrices in stage one, with similar risk estimates and boundaries obtained as those in Section 6.6.

In the analysis of this chapter I do not consider any covariates, but the approach can be used in ecology studies to explore the risk factors that might explain the spatio-temporal variation in respiratory disease risk, such as environment exposures, smoking, socio-economic deprivation, etc. Furthermore, the graph-based optimisation algorithm (Lee et al., 2021) used to generate candidate neighbourhood matrices has limitations. Firstly, the current implementation makes use of a local search method that is not guaranteed to find the global optimal matrix. Secondly, the running-time of the algorithm depends exponentially on the number of edges, which could be very high when dealing with a large number of edges. Therefore, care should be taken when choosing the value of M, i.e. the number of candidate matrices to be generated in stage one. If M is too small then the best neighbourhood matrix may not exist in the candidates, whereas if M is too large then the computational burden could be huge. An optimisation algorithm with provable guarantees of a global optimum solution and higher efficiency can be investigated in more details in a future study. If a global optimal neighbourhood matrix can be found, then fitting a model based on this global optimum would be much simpler and faster than the method proposed here that estimates an appropriate value from a set of candidates based on many local optima, although it does not allow to make probability statements such as calculating the probability of each edge being a boundary. Other avenues for future work include extending the proposed model from count data to model Gaussian or binomial type data, allowing the boundaries in disease risk to evolve dynamically over time, and extending the spatio-temporal model to the multivariate domain to consider multiple diseases simultaneously.

Chapter 7

Conclusion

This thesis focused on estimating the spatial and spatio-temporal patterns in disease risk, and identifying discontinuities in the risk surface. In most disease mapping studies that assess the extent and pattern of disease risk, the study region is split into non-overlapping areal units and then disease risk is estimated for each of these areal units. Disease maps can be used to quantify the spatial inequalities in ill health and provide a visual representation of the risk patterns for policy makers, by shading the areal units in different colours based on their disease risks. The standardised incidence ratio (SIR) is an unstable estimate of disease risk when the population at risk is small or the disease in question is rare. Therefore, model-based approaches to the analysis of disease maps are beneficial. Bayesian hierarchical modelling is mostly used to estimate disease risk patterns by utilising a Poisson log-linear generalised linear mixed model, which includes known covariate risk factors that impact on disease risk and a set of random effects that account for the spatial autocorrelation present in the disease data. The random effects are most commonly modelled by conditional autoregressive (CAR) models introduced in Section 2.4.4, which assume spatial autocorrelation between pairs of geographically neighbouring areal units by typically inducing a neighbourhood matrix defined using the border sharing rule. These CAR models smooth the random effects (or disease risks) in geographically adjacent areas towards each other. However in practice, real spatial data are likely to contain areas of smooth evolution in disease risk as well as neighbouring areas that exhibit substantially different disease risks. Thus enforcing a constant level of spatial smoothness across the entire spatial region may obscure any discontinuities in the disease risk surface, hinder the identification of high-risk areas and produce poor risk estimates. As a consequence, there has been growing interest in developing modelling approaches which allow for discontinuities in the spatial risk pattern. Two fields are related to this, namely spatial clustering and boundary analysis (see Section 2.6). The former allows for discontinuities by identifying clusters of areas that exhibit an elevated risk of disease compared to their surrounding areas, while the latter allows for discontinuities by detecting the locations of boundaries where there are large step-changes in disease risk between geographically adjacent areal units. In this thesis I developed spatial and spatio-temporal methodology that can carry out both risk estimation and discontinuity identification.

7.1 Spatial clustering approaches

Chapter 3 outlined a spatial modelling approach for estimating the spatial pattern in disease risk and identifying clusters of high (or low) risk areas. In the first stage the approach uses k-means clustering to create a set of candidate cluster structures, each of which is used to construct a candidate neighbourhood matrix. In the second stage separate Bayesian hierarchical models are applied to the data for each candidate neighbourhood matrix. The most appropriate neighbourhood matrix, which corresponds to the most appropriate cluster structure, is chosen using model selection rules such as the Deviance Information Criterion and the effective number of independent parameters. A simulation study has shown that the model with the cluster structure minimising the effective number of parameters exhibited consistently good performance in terms of both risk estimation and cluster identification, and particularly performed better than the Leroux model (Leroux et al., 2000). This modelling approach treats the identification of the cluster structure as a model comparison problem. It always identifies a single optimal cluster structure for the data, which is straightforward to implement and understand for non-specialist users. However the approach does not quantify the uncertainty in the cluster structure, because the neighbourhood matrix is fixed during the modelling procedure when fitting each model.

Chapter 4 introduced an alternative spatial model which allows for uncertainty in the cluster structure when estimating disease risk in the second stage. Here a variety of clustering methods are used in stage one to construct a much bigger set of candidate cluster structures than that in Chapter 3, including k-means clustering, k-medoids clustering, hierarchical agglomerative clustering with centroid, complete, average and Ward linkage, divisive clustering and expectation-maximisation clustering. The near 8-fold increase in the set of candidate cluster structures gives much greater flexibility in cluster identification. Likewise, each potential cluster structure is used to generate a candidate neighbourhood matrix. In stage two, a single Bayesian spatial model is fitted to jointly estimate the disease risk across the study region and an appropriate cluster/discontinuity structure. The latter is achieved by directly modelling the spatial correlation structure in the random effects, where the neighbourhood matrix $\widetilde{\boldsymbol{W}}$ is treated as a parameter to be estimated from the set of candidates generated in stage one. This methodology does not require a comparison of multiple models, which substantially reduces the computational time. In addition the model may select a different neighbourhood matrix at each iteration of MCMC simulation, thus the uncertainty in the cluster structure can be quantified and propagated through the model. Two approaches were proposed for updating the choice of $\widetilde{\boldsymbol{W}}$ in MCMC simulation. **Approach 1** updates $\widetilde{\boldsymbol{W}}$ by using a Metropolis-Hastings step consisting of two MCMC moves. Specifically, the first move uniformly draws a new proposal matrix from the candidate neighbourhood matrices whose corresponding cluster structures are generated from the same clustering method as the current matrix but with a different number of clusters, while the second move proposes a new value from the candidate matrices whose corresponding cluster structures have the same number of clusters as the current choice but are generated from a different clustering method. In contrast, **Approach 2** updates $\widetilde{\boldsymbol{W}}$ via a Metropolis-Hastings step which only proposes a new neighbourhood matrix once at each MCMC iteration. The simulation study presented in Section 4.3 showed that the first approach performed better than the second approach due to it providing slightly more accurate risk estimates and less varied results. The study also illustrated that in the presence of clusters and a relatively high number of expected cases (i.e. greater than 30), the proposed model overall exhibited improved performance for estimating risks compared to the Leroux model, and also behaved well in identifying accurate cluster structures with high adjusted Rand Index (ARI) values being obtained in most cases. This is because the proposed model accounts for the clusters in risk by estimating an appropriate neighbourhood matrix, which better represents the spatial autocorrelation structure in the data.

The approach outlined in Chapter 4 is in a purely spatial setting, which was then extended to the spatio-temporal domain in Chapter 5. Here the spatio-temporal variation in the data is decomposed into an overall temporal trend common to all areal units and separate spatial surfaces for each time period. The temporal trend is modelled by a first order autoregressive process, while the spatial surface is modelled by a separate Leroux CAR prior. Two different model variants are considered, where the spatial clusters either remain fixed or evolve over time. In model **ST-A*** the spatial clusters remain fixed during the entire study period, thus a neighbourhood matrix $\widetilde{\boldsymbol{W}}$ common to all time periods is estimated.

In model **ST-B*** the clusters are allowed to vary dynamically over time, thus a separate neighbourhood matrix $\widetilde{\boldsymbol{W}}_t$ is estimated for each time period. A simulation study was carried out to comprehensively assess the performance of the cluster models and compare them to an existing non-clustering model proposed by Napier et al. (2016). The proposed models provided better risk estimates than the Napier et al. (2016) model in the cases where there are clusters present in the data, and also identified accurate clusters as measured by the adjusted Rand Index. When the risk surface is spatially smooth for each time period, model **ST-A*** performed as well as the Napier et al. (2016) model. The cluster models performed less well when the clusters are not pronounced and the expected disease counts are very small, because in this scenario the clusters are hardest to identify based on their small size and small numbers of disease cases. Comparing models **ST-A*** and **ST-B***, the former with constant clusters over time is more appropriate if the disease data have a high correlation in time, whereas if the disease data are less temporally correlated and the cluster structures in particular years are of interest, then the latter with temporally evolving clusters is a better choice. In addition, in Chapter 4 the spatial dependence parameter ρ is fixed at 0.99 to enforce strong spatial autocorrelation globally, so that the spatial correlation structure can be modelled locally by estimating an appropriate neighbourhood matrix for the data. However, whether we should fix the spatial dependence parameter or estimate it in the model is a good question that should be considered. Therefore, in Chapter 5 I compared the modelling performance when fixing the global spatial dependence parameters (ρ_s, ρ_{s_t}) at 0.99 and also when estimating them within the model. The simulation study showed that estimating (ρ_s, ρ_{s_t}) produced better results overall than fixing them at 0.99, although the latter needed less computational time. Based on this result, I recommend estimating (ρ_s, ρ_{s_t}) as part of the model rather than fixing them when estimating disease risk.

The approaches described above have a common characteristic of accounting for clusters via the spatial autocorrelation structure of the random effects. These approaches estimate the disease risk and the cluster structure by only allowing for correlation between neighbouring areal units that lie in the same cluster. If two adjacent areas are in different clusters then spatial autocorrelation is not enforced between them and thus the estimated risks in these areas are not smoothed towards each other. Furthermore, the clusters identified by our approaches represent the number of distinct risk levels rather than the number of spatially contiguous clusters. This is because each clustering method is applied to the data without regard to the spatial positions of the areal units when generating the candidate cluster structures. Therefore, the identified clusters can contain areal units which are far apart geographically but exhibit similar disease risks during the study period.

7.2 Boundary analysis

Unlike the approaches in Chapters 3, 4 and 5 that allow for risk discontinuities by seeking for clusters of areas with similar or different disease risks, Chapter 6 introduced an approach focusing on identifying the boundaries in the risk surface that separate geographically adjacent areas that have very different risks. This two-stage approach is an extension of Lee et al. (2021) by allowing for variation in the neighbourhood matrix, and hence in the boundaries identified, in the modelling process. In stage 1, a graph-based optimisation algorithm is used to estimate multiple candidate neighbourhood matrices, which represent a range of possible boundary structures in the data. The algorithm views the areal units as the vertices of a graph and the neighbour relations as the set of edges. It estimates whether each edge in the graph should be removed or not based on an objective function. If the edge between adjacent areas (i, j) has been removed, then the value of the element w_{ij} in the border sharing $\boldsymbol{W}$ is changed from 1 to 0, suggesting the presence of a boundary between the two areas. In stage 2, a Bayesian model is fitted to the data, where the spatio-temporal risk pattern is represented with a multivariate first order autoregressive process with a spatially correlated precision matrix. The neighbourhood matrix $\widetilde{\boldsymbol{W}}$, representing a boundary structure for the data, is estimated from the set of candidates constructed in stage 1 when estimating disease risk, rather than naively being fixed based on geographical adjacency. The simulation study in Section 6.4 showed that the proposed model exhibited much better risk estimation and model fit than the global smoothing model developed by Rushworth et al. (2014) in the presence of boundaries/discontinuities. This improved estimation is because our model can flexibly capture either spatial smoothness or a boundary of step change in the data between adjacent areas, while the global smoothing model enforces inappropriate spatial smoothing between these areas that have very different data values, which leads to less accurate risk estimates. The proposed model also performed well in terms of detecting the true locations of boundaries in simulated data as measured by the AUC (area under the curve) statistics. In addition, the study showed that the proposed model and the model developed by Lee et al. (2021) performed broadly similarly in terms of risk estimation. This is probably because the candidate neighbourhood matrices obtained in stage one do not differ much from each other, thus the contribution of the model allowing for uncertainty in $\widetilde{\boldsymbol{W}}$ to risk estimation is not very pronounced.

7.3 Results from applications to the Greater Glasgow respiratory disease data

Both of the spatial models introduced in Chapters 3 and 4 were applied to respiratory disease data for the Greater Glasgow and Clyde Health Board region for 2016 to estimate the spatial pattern in disease risk and identify possible clusters of areas of higher and lower risks. Model P4 in Chapter 3 identified a cluster structure containing 5 distinct clusters or risk levels, which produces the lowest effective number of parameters, while modelling **Approach 1** in Chapter 4 favoured a cluster structure with 2 clusters, which corresponds to the posterior mode of $\widetilde{\boldsymbol{W}}$. Both models picked out the high-risk clusters in the east and north of Glasgow city center such as Easterhouse, Drumchapel, Springburn and Possilpark. They also identified some high risk areas to the south of the Clyde river, containing Govan, Nitshill and Priesthill. Conversely, the low-risk clusters were found in the north east of the city (e.g. Bearsden and Lennoxtown), in the prosperous West End of Glasgow (e.g. Kelvinside and Jordanhill), as well as to the extreme south (e.g. Clarkston, Newton Mearns and Eaglesham). In addition, model P4 also detected a cluster of areas with a medium level of disease risk, for example the large rural areas of Inverclyde in the far west. These results suggested that people living in deprived areas tend to have higher respiratory risks than those living in affluent areas.

The spatio-temporal models proposed in Chapter 5 were applied to respiratory hospital admission data in the Greater Glasgow and Clyde Health Board from 2011 to 2017, with the aim of estimating the spatial disease risk pattern over time and identifying the, possibly temporally evolving, cluster/discontinuity structures in disease risk. Model **ST-A*** assumes that the spatial clusters are constant over time, whereas model **ST-B*** allows them to evolve over time. The two models produced similar disease risk patterns and showed a generally increasing trend in risk over the 7-year period, with the average disease risk increasing by 16%. Model **ST-A*** selected a cluster structure with 2 distinct clusters (or risk levels) and this structure is common to all time periods. In contrast, model **ST-B*** detected 5 or 6 different cluster levels depending on the year. The two models identified a number of the same areas that have a high level of disease risk, and many of them are the same areas that were identified as having high risks in 2016 in Chapters 3 and 4, such as Easterhouse, Govan, Drumchapel and Summerston. They also captured the locations of cluster discontinuities, the majority of which occurred between neighbourhoods that exhibited very different disease risks. For example, the high-risk areas Drumchapel and Drumry to the north west of the health board are bordered to the north by the more affluent and low-risk Bearsden area. The single IZ in the north east of the city near Kirkintilloch had an elevated risk compared to its geographical neighbours. Model **ST-B*** captured the evolution of clusters over time. For example, areas of Inverclyde were in a low-risk cluster in 2011, whereas by 2017 they joined a moderately high-risk cluster.

The methodology presented in Chapter 6 was applied to the same respiratory disease data as in Chapter 5. This approach produced broadly similar risk surfaces to those estimated from Chapter 5, and also provided additional insight as to the locations of boundaries corresponding to sizeable changes in disease risk between adjacent IZs. The model identified 277 boundaries in the spatial risk pattern common to all time periods, which comprise 41% of the total number of edges in the study region. The largest boundaries were identified between Drumchapel and its neighbour Bearsden in the north-west, as well as between Easterhouse and Garrowhill in the East End of the city. The results also showed that some identified boundaries were close to physical barriers e.g. motorways, railways or big roads, which are difficult to cross and make people living on either side hard to mix.

7.4 Limitations and future work

This thesis provided new approaches for estimating the disease risk and identifying discontinuities in the spatial risk pattern in both spatial and spatio-temporal data. The approaches proposed in Chapters 3, 4 and 5 identify clusters of areas that have elevated or reduced risks compared to their neighbours. Here the clusters identified represent the number of distinct disease risk levels and so are not spatially contiguous. However, it is straightforward to induce spatial contiguity in the sets of clusters by a post-processing step that simply relabels the non-contiguous parts as new clusters. Alternatively, the clustering methods used in the first stage can be adapted to respect the spatial contiguity structure of the study region. Note that the clustering approaches proposed here require the user to define the maximum number of clusters (risk levels) appropriately when generating the candidate cluster structures. One limitation to the spatio-temporal model with temporally evolving clusters in Chapter 5 is that it might pick out some erroneous discontinuities, which is because each candidate cluster structure is elicited using data for a single year and thus could be affected by random noise in the observed disease counts. Therefore future work could consider estimating the candidate cluster structures for a given year by clustering the data for the year in question and the q years before and after. The spatio-temporal models in Chapters 5 and 6 were applied to the data collected for only 7 time points (2011-2017). If the data could be available for a longer time period, more complex modelling of the temporal trend could be considered. The analysis of this thesis only focuses on a single disease, thus it would be of interest to extend the proposed models to spatial or spatio-temporal multivariate disease models. Models based on multiple diseases allow for a more comprehensive and better understanding of the health profiles in Greater Glasgow. Such a model could be $\ln(R_{itd}) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \phi_{itd} + \theta_{td}$. Here R_{itd} represents the disease risk in areal unit i during time period t for disease d, which is modelled by a disease specific space-time effect ϕ_{itd} and a disease specific temporal effect θ_{td}. The former can account for spatial and between disease correlation in data by a multivariate space-time CAR model proposed by Gelfand and Vounatsou (2003), while the latter can account for temporal autocorrelation by a first (AR(1)) or a second (AR(2)) order autoregressive process.

The proposed models in this thesis can also be applied to an ecological regression analysis which estimates the health impact of both good and bad exposures, by including covariates that are thought to be relevant in the explanation of disease risk variation over space and time, such as socio-economic deprivation, demography, air pollution concentrations and greenspace. The methods proposed in this thesis are based on aggregated disease count data at the areal unit level, and they assume that the level of disease risk is constant within each areal unit. Such methods are preferable in certain contexts, for instance, when disease interventions and public health policies are employed on small areas and thus public health experts and stakeholders are interested in areal-based estimates. However, in ecological studies the aggregation of data causes the loss or concealment of certain details about individuals. This may result in the ecological fallacy (Wakefield and Salway, 2001), which occurs when an inference at the individual level is made simply based on aggregated data for a group that those individuals belong to. Furthermore, the unknown spatial confounding is likely to be driven by quantities that vary continuously in space (e.g. temperature, air pollution, etc). Therefore, another area for future work could be to address the problems by developing continuous domain models for disease mapping. Such an approach could create a set of grid squares over the study region and then transform the areal unit data to the grid level scale, so that we can make continuous inference in disease risk at the fine grid level. When the grid squares become smaller, the inference will get closer to an individual level and the disease risk on a spatially continuous risk pattern will be estimated.

Bibliography

Adin, A., Lee, D., Goicoa, T. and 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.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. and Ronquist, F. (2004), 'Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference', *Bioinformatics* **20**(3), 407–415.

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

Anderson, C., Lee, D. and Dean, N. (2016), 'Bayesian cluster detection via adjacency modelling', *Spatial and spatio-temporal epidemiology* **16**, 11–20.

Anderson, C., Lee, D. and Dean, N. (2017), 'Spatial clustering of average risks and risk trends in Bayesian disease mapping', *Biometrical Journal* **59**(1), 41–56.

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

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

Besag, J., York, J. and Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the institute of statistical mathematics* **43**(1), 1–20.

Bradley, A. P. (1997), 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', *Pattern recognition* **30**(7), 1145–1159.

Brooks, S. (1998), 'Markov chain Monte Carlo method and its application', *Journal of the royal statistical society: series D (the Statistician)* **47**(1), 69–100.

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

Chiang, C. L., World Health Organization et al. (1979), 'Life table and mortality analysis'.

Chung, F., Barnes, N., Allen, M., Angus, R., Corris, P., Knox, A., Miles, J., Morice, A., O'Reilly, J. and Richardson, M. (2002), 'Assessing the burden of respiratory disease in the UK', *Respiratory medicine* **96**(12), 963–975.

Collinson, P. (1998), 'Of bombers, radiologists, and cardiologists: time to ROC', *Heart* **80**(3), 215–217.

De Boor, C. (1972), 'On calculating with B-splines', *Journal of Approximation theory* **6**(1), 50–62.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

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

Earl, D. J. and Deem, M. W. (2005), 'Parallel tempering: Theory, applications, and new perspectives', *Physical Chemistry Chemical Physics* **7**(23), 3910–3916.

Eddelbuettel, D. (2013), *Seamless R and C++ integration with Rcpp*, Springer.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J. and Bates, D. (2011), 'Rcpp: Seamless R and C++ integration', *Journal of Statistical Software* **40**(8), 1–18.

Elliot, P., Wakefield, J. C., Best, N. G., Briggs, D. J. et al. (2000), *Spatial epidemiology: methods and applications.*, Oxford University Press.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H. and Zubrzycki, S. (1951), Sur la liaison et la division des points d'un ensemble fini, *in* 'Colloquium mathematicum', Vol. 2, pp. 282–285.

Forum of International Respiratory Societies (2017), *The global impact of respiratory disease*, European Respiratory Society.

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

Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010), *Handbook of spatial statistics*, CRC press.

Gelfand, A. E. and Vounatsou, P. (2003), 'Proper multivariate conditional autoregressive models for spatial data analysis', *Biostatistics* **4**(1), 11–15.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995), *Bayesian data analysis*, Chapman and Hall/CRC.

Gelman, A., Roberts, G. O., Gilks, W. R. et al. (1996), 'Efficient Metropolis jumping rules', *Bayesian statistics* **5**(599-608), 42.

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

Geman, S. and Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Geweke, J. (1992), 'Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments', *Bayesian statistics* **4**, 641–649.

Glasgow City Council (2020), Child Poverty in Glasgow Report 2020, Technical report, Centre for Civic Innovation.

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

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (ROC) curve.', *Radiology* **143**(1), 29–36.

Hartigan, J. A. (1981), 'Consistency of single linkage for high-density clusters', *Journal of the American Statistical Association* **76**(374), 388–394.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer, New York.

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

Hubert, L. and Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**(1), 193–218.

Jacquez, G. M., Maruca, S. and Fortin, M.-J. (2000), 'From fields to objects: a review of geographic boundary analysis', *Journal of Geographical Systems* **2**(3), 221–241.

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.

Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. M. (1998), 'Markov chain Monte Carlo in practice: a roundtable discussion', *The American Statistician* **52**(2), 93–100.

Kaufman, L. and Rousseeuw, P. J. (2009), *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons.

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. and Besag, J. (1998), 'Modelling risk from a disease in time and space', *Statistics in medicine* **17**(18), 2045–2060.

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

Kofke, D. A. (2002), 'On the acceptance probability of replica-exchange Monte Carlo trials', *The Journal of chemical physics* **117**(15), 6911–6914.

Kuiper, F. K. and Fisher, L. (1975), '391: A Monte Carlo comparison of six clustering procedures', *Biometrics* **31**(3), 777–783.

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

Kulldorff, M., Huang, L. and Konty, K. (2009), 'A scan statistic for continuous data based on the normal probability model', *International journal of health geographics* **8**(1), 1–9.

Lambert, D. (1992), 'Zero-inflated Poisson regression, with an application to defects in manufacturing', *Technometrics* **34**(1), 1–14.

Lawson, A. B., Banerjee, S., Haining, R. P. and Ugarte, M. D. (2016), *Handbook of spatial epidemiology*, CRC Press.

Lawson, A. and Lee, D. (2017), Bayesian disease mapping for public health, *in* ‘Handbook of statistics’, Vol. 36, Elsevier, pp. 443–481.

Lee, D. (2011), ‘A comparison of conditional autoregressive models used in Bayesian disease mapping’, *Spatial and spatio-temporal epidemiology* **2**(2), 79–89.

Lee, D., Meeks, K. and Pettersson, W. (2021), ‘Improved inference for areal unit count data using graph-based optimisation’, *Statistics and Computing* **31**(4), 1–17.

Lee, D. and Mitchell, R. (2012), ‘Boundary detection in disease mapping studies’, *Biostatistics* **13**(3), 415–426.

Lee, D. and Mitchell, R. (2013), ‘Locally adaptive spatial smoothing using conditional autoregressive models’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(4), 593–608.

Lee, D., Rushworth, A. and Napier, G. (2018), ‘Spatio-temporal areal unit modelling in R with conditional autoregressive priors using the CARBayesST package’, *Journal of Statistical Software* **84**(9).

Lee, D., Rushworth, A. and Sahu, S. K. (2014), ‘A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution’, *Biometrics* **70**(2), 419–429.

Leroux, B. G., Lei, X. and Breslow, N. (2000), Estimation of disease rates in small areas: a new mixed model for spatial dependence, *in* ‘Statistical models in epidemiology, the environment, and clinical trials’, Springer, pp. 179–191.

Li, P., Banerjee, S. and McBean, A. M. (2011), ‘Mining boundary effects in areally referenced spatial data using the Bayesian information criterion’, *Geoinformatica* **15**(3), 435–454.

Link, W. A. and Eaton, M. J. (2012), ‘On thinning of chains in MCMC’, *Methods in ecology and evolution* **3**(1), 112–115.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S. Y. et al. (2012), ‘Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010’, *The lancet* **380**(9859), 2095–2128.

Lu, H. and Carlin, B. P. (2005), 'Bayesian areal wombling for geographical boundary analysis', *Geographical Analysis* **37**(3), 265–285.

Lu, H., Reilly, C. S., Banerjee, S. and Carlin, B. P. (2007), 'Bayesian areal wombling via adjacency modeling', *Environmental and ecological statistics* **14**(4), 433–452.

Ma, H., Carlin, B. P. and Banerjee, S. (2010), 'Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis', *Biometrics* **66**(2), 355–364.

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

MacNab, Y. C. and Gustafson, P. (2007), 'Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance', *Statistics in medicine* **26**(24), 4455–4474.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297.

Madhulatha, T. S. (2011), Comparison between k-means and k-medoids clustering algorithms, *in* 'International Conference on Advances in Computing and Information Technology', Springer, pp. 472–481.

Madhulatha, T. S. (2012), 'An overview on clustering methods', *arXiv preprint arXiv:1205.1117* .

McCartney, G. (2012), What would be sufficient to reduce health inequalities in Scotland?, Report to the ministerial taskforce on health inequalities, NHS Health Scotland.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models, Second Edition*, Chapman & Hall, London.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.

Mitchell, R. and Lee, D. (2014), 'Is there really a "wrong side of the tracks" in urban areas and does it matter for spatial analysis?', *Annals of the Association of American Geographers* **104**(3), 432–443.

Moore, E. H. (1920), 'On the reciprocal of the general algebraic matrix', *Bulletin of the American Mathematical Society* **26**, 394–395.

Moran, P. A. (1950), 'Notes on continuous stochastic phenomena', *Biometrika* **37**(1/2), 17–23.

Murtagh, F. and Legendre, P. (2014), 'Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?', *Journal of classification* **31**(3), 274–295.

Napier, G., Lee, D., Robertson, C. and 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. and 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. and Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

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

Park, H.-S. and Jun, C.-H. (2009), 'A simple and fast algorithm for K-medoids clustering', *Expert systems with applications* **36**(2), 3336–3341.

Penrose, R. (1955), A generalized inverse for matrices, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 51, Cambridge University Press, pp. 406–413.

R Core Team (2013), *R: A language and environment for statistical computing*, Vienna, Austria.

Robbins, H. E. (1992), An empirical Bayes approach to statistics, *in* 'Breakthroughs in statistics', Springer, pp. 388–394.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Rushworth, A., Lee, D. and 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.

Rushworth, A., Lee, D. and Sarran, C. (2017), 'An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**(1), 141–157.

Salciccioli, J. D., Marshall, D. C., Shalhoub, J., Maruthappu, M., De Carlo, G. and Chung, K. F. (2018), 'Respiratory disease mortality in the United Kingdom compared with EU15+ countries in 1985-2015: observational study', *British Medical Journal* **363**.

Santafé, G., Adin, A., Lee, D. and 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. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Seaman, V. (1798), *An inquiry into the cause of the prevalence of the yellow fever in New-York*, T. and J. Swords.

Smith, B. J. (2005), 'Bayesian output analysis program (BOA) version 1.1 user's manual', *Dept. of Biostatistics, Univ. of Iowa, College of Public Health, http://www. public-health. uiowa. edu/boa* .

Sneath, P. H. (1957), 'The application of computers to taxonomy', *Microbiology* **17**(1), 201–226.

Sneath, P. H., Sokal, R. R. et al. (1973), *Numerical taxonomy. The principles and practice of numerical classification*, W.H. Freeman and Company, San Francisco.

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

Sokal, R. R. and Michener, C. D. (1958), 'A statistical method for evaluating systematic relationships.', *The University of Kansas science bulletin* **38**, 1409–1438.

Sorensen, T. A. (1948), 'A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons', *Biol. Skar.* **5**, 1–34.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and 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.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996), 'BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii)', *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK* pp. 1–59.

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

Takahashi, K., Kulldorff, M., Tango, T. and Yih, K. (2008), 'A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring', *International Journal of Health Geographics* **7**(1), 1–14.

The Glasgow Centre for Population Health (2021), 'Understanding Glasgow-The Glasgow Indicators Project. Life expectancy comparisons-UK cities', `https://www.understandingglasgow.com/indicators/health/comparisons/uk_cities`. Accessed: 2021.

Tibshirani, R., Walther, G. and Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423.

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

Torabi, M. and Rosychuk, R. J. (2011), 'Spatio-temporal modelling using B-spline for disease mapping: analysis of childhood cancer trends', *Journal of Applied Statistics* **38**(9), 1769–1781.

Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R., Albertson, S. B., Deshpande, A., Farag, T., Abebe, Z. et al. (2018), 'Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016', *The Lancet infectious diseases* **18**(11), 1191–1210.

Ugarte, M., Etxeberria, J., Goicoa, T. and Ardanaz, E. (2012), 'Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990–2005)', *Cancer Epidemiology* **36**(3), 254–262.

Ugarte, M., Goicoa, T. and 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. and Kim, A. (2013), ‘A Bayesian model for cluster detection’, *Biostatistics* **14**(4), 752–765.

Wakefield, J. and Salway, R. (2001), ‘A statistical framework for ecological and aggregate studies’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **164**(1), 119–137.

Waller, L. A. and Carlin, B. P. (2010), ‘Disease mapping’, *Chapman & Hall/CRC handbooks of modern statistical methods* **2010**, 217.

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

Walsh, D., Bendel, N., Jones, R. and Hanlon, P. (2010), Investigating a Glasgow Effect: Why do equally deprived UK cities experience different health outcomes?, Technical report, Glasgow Centre for Population Health.

Ward Jr, J. H. (1963), ‘Hierarchical grouping to optimize an objective function’, *Journal of the American statistical association* **58**(301), 236–244.

Yim, O. and Ramdeen, K. T. (2015), ‘Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data’, *The quantitative methods for psychology* **11**(1), 8–21.