Pseudo-continuous spatial and spatio-temporal modelling of disease risk

Kamol Sanittham

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

March 2021

Abstract

Disease mapping approach is the statistical methodology used to estimate disease risk over time, and is generally based on areal unit data. Conditional autoregressive (CAR) models are the most common modelling approach for disease data at the areal unit level. Such approaches assume constant disease risk within each areal unit, which may not be realistic. Therefore this study aims to address this problem by creating a pseudo continuous disease risk surface over the Greater Glasgow and Clyde Health Board. A set of regular grid squares is overlaid across the study region and the main focus of this study is to estimate disease risk in each grid square after removing grid squares with zero population. Areal unit data are transformed to the grid level via two novel approaches which are multiple imputation and data augmentation and then use these grid data to fit the standard Leroux CAR model to estimate the spatial patterns in disease risk at the grid level. The multiple imputation approach generates multiple sets of disease counts at the grid level via multinomial sampling, and each dataset is used to fit the CAR model then combine the results to estimate the grid level disease risk. While the data augmentation allows uncertainty in the disease counts by updating them in the MCMC steps. Each method is applied to respiratory hospital admission data from the Greater Glasgow and Clyde Health Board area. The final piece of work of this thesis extends the spatial model to measure health inequality in Glasgow over time. Overall, it was found that disease risk is increasing over time and the areas with higher risk correspond to the deprived areas, while areas with lower risk tend to be the wealthier areas in Glasgow.

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors Prof Duncan Lee and Dr Craig Anderson for giving me the opportunity for this PhD study, and providing invaluable guidance throughout this study. Their vision, motivation, patience, and enormous knowledge have extremely inspired me. This accomplishment would not have been possible without them. I am also thankful to the examiners, Dr Susanna Cramb (Queensland University of Technology) and Dr Mayetri Gupta (University of Glasgow), for valuable suggestions that make my thesis a better version.

I gratefully acknowledge the funding received towards my PhD from Ministry of Higher Education, Science, Research and Innovation, Royal Thai Government. I am also thankful to Chiangmai Rajabhat university for allowing me this greatest opportunity in life. I very much appreciate to all my teachers for all level of my education. In particular, Asst Prof Yuwanit Hongtrakul and Assoc Prof Putipong Bookamana for their guidance and support, not only in academic perspective but also living aspect.

I would also like to extend my sincere thanks to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. I would not be in the position I am today without them. To my friends, thank you so much for supporting me throughout my PhD journey, for badminton, travels, concerts, and parties. I could not imagine my life in Glasgow without you guys. Thank you.

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at University of Glasgow, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree. I have acknowledged all sources used and have cited these in the bibliography section.

Further, I delivered an invited talk on this work at the GEOMED conference in Glasgow UK, in 2019.

Chapter 1

Introduction

Disease risk varies in space and time, and is often related to social factors such as smoking, drinking, diet and environmental factors such as air pollution or water quality (Lawson and Lee, 2017). Poverty is also one of the major factors that can be observed in the variation in disease risks, where the areas with higher disease risk are more likely to be deprived areas. In contrast, lower risks are often related to wealthier areas (McCartney, 2012). This difference in disease risk across different social groups is unfair and defined as health inequality by NHS Health Scotland (2015). Therefore government and health authorities take a huge interest to improve their people's health as a whole by increasing life expectancy and reducing the health inequalities gap (Oliver, 2001).

In order to explore spatial variation in disease risks, one of the first studies by Seaman (1798) showed the spread of yellow fever in New York in 1798 by producing a dot disease map shown in Figure 1.1. In 1854, there was a cholera outbreak in Soho, London. At the time, it was assumed that cholera was transmitted via air, but Snow (1855) produced the disease map illustrated in Figure 1.2, which shows the locations of the disease cases. He identified that the cholera cases were found near a water pump in Broad (now Broadwick) street. Therefore, the water pumps were eventually removed and modernisation of water supplies and sanitation systems was carried out in London. Current studies also illustrate the spatial patterns of such health inequalities via disease maps where areas are shaded on a scale in different colours which displays the disease risk (MacNab et al., 2004; Wakefield, 2007; Lee, 2011; Rushworth et al., 2014). An

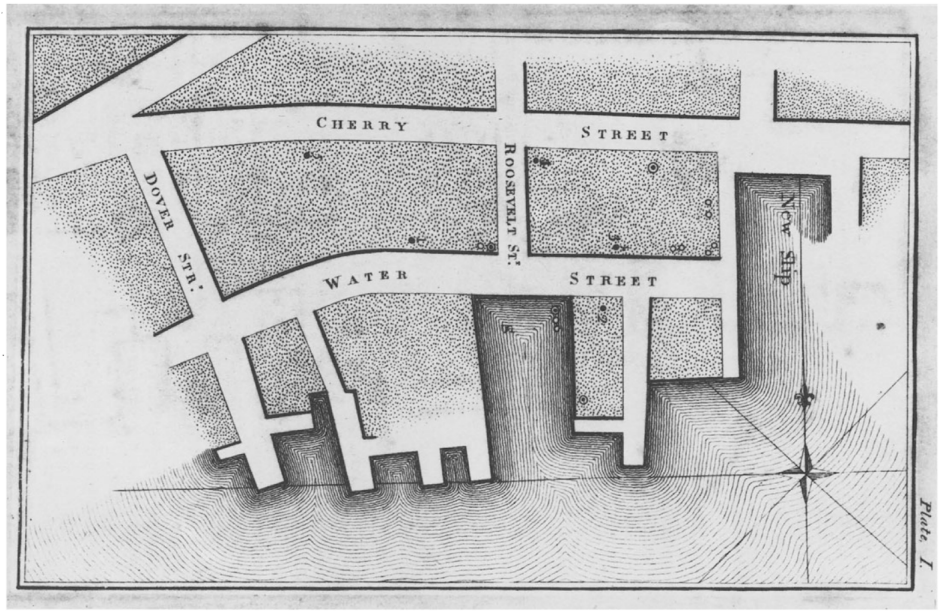


Figure 1.1: One of the first dot maps produced by Seaman (1798).

example of a disease risk map is presented in Figure 1.3, which related to the estimates of respiratory disease in the Greater Glasgow and Clyde Health Board for the years 2015 - 2016.

The data used to create such disease maps are collected from non-overlapping areal units, such as electoral wards, census tracts, or health board areas. Data at the individual level can not be made publicly available due to confidentiality reasons, therefore only aggregated data in each area are available. Each areal unit normally has different population sizes and demographic structures (e.g. age and sex profiles), so the comparison of disease risk between each areal unit can usually be made via a standardised incidence ratio (SIR). The SIR can be calculated as the ratio of the observed number of disease cases to the expected number of disease cases. The expected number of disease case in each area is computed via indirect standardisation based on its population demographics. An SIR value less than 1 indicates that there are fewer disease cases than expected, and hence lower disease risk, while an SIR greater than 1 indicates that there are more disease cases than expected, hence higher disease risk. More specifically, an SIR value of 0.9 corresponds to a disease risk which is 10% lower than expected, while an SIR value of 1.1 corresponds to a disease risk which is 10% higher than expected. However, in cases where the population of the study is small, or the

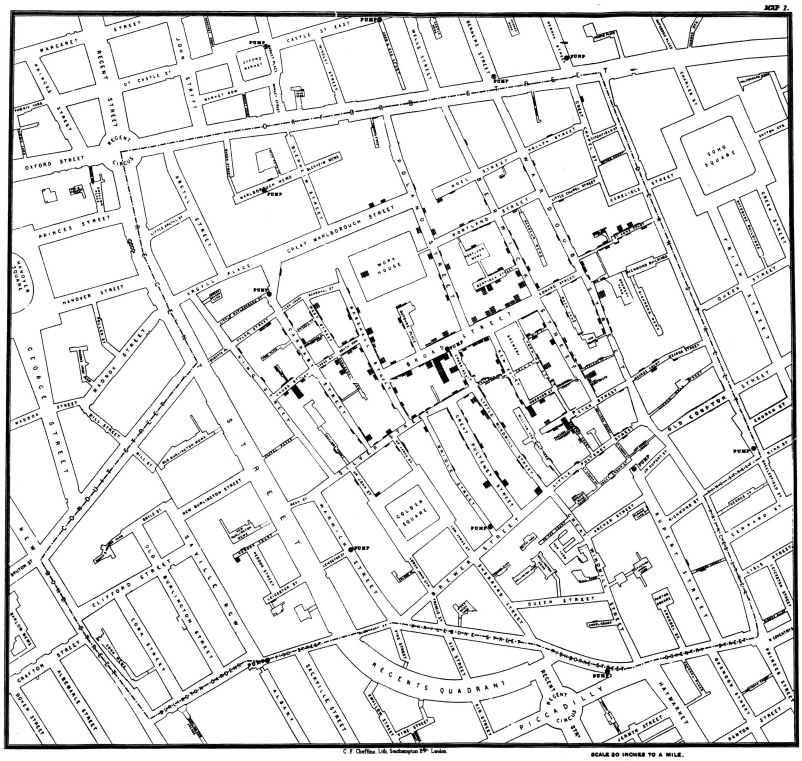


Figure 1.2: Disease map of cholera in Soho, London (Snow, 1855)

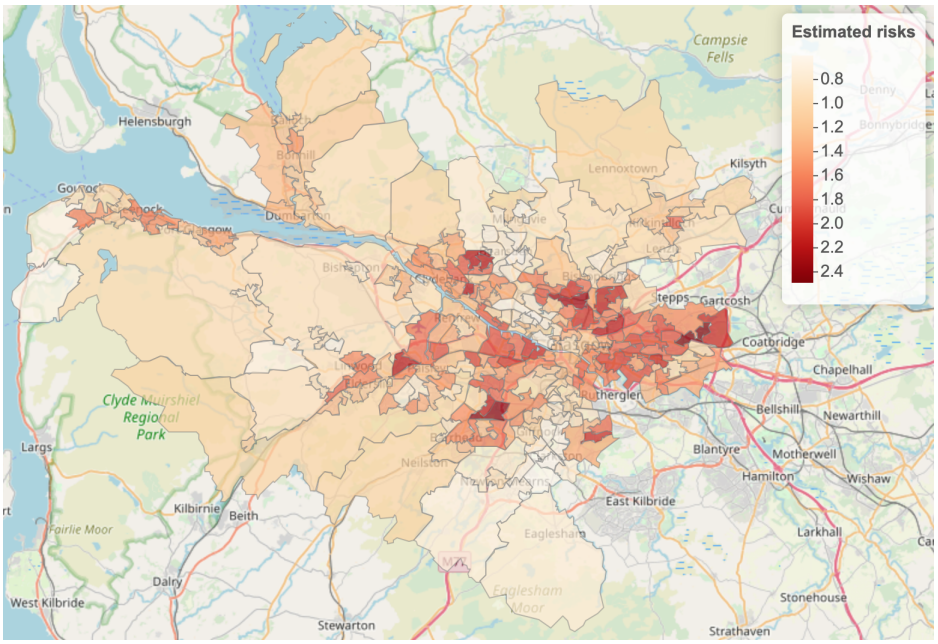


Figure 1.3: An example of disease mapping for respiratory disease in Glasgow in the years 2015 - 2016.

disease being studied is rare, some areas may have low values of the expected number of disease cases. This can lead to the SIR giving an unstable estimate of disease risk.

In order to overcome this limitation, hierarchical Bayesian modelling is commonly adopted to estimate disease risk patterns by utilising a Poisson generalised linear model with random effects that account for the spatial autocorrelation present in the data. These random effects are typically modelled via a conditional autoregressive (CAR) models (Besag et al., 1991) which are outlined in Chapter 2. These models make an assumption of spatial autocorrelation; i.e. nearby areas are more likely to have similar disease risks than areas further apart. This is because of the idea that the neighbouring areas are more likely to have similar social characteristics, e.g. ethnicity, house price, population behaviour, etc. These models also address the problem of overdispersion that often present in count data.

An example related to these models is presented by Lee and Mitchell (2012), they utilised this idea to estimate risk surface for lung cancer in the Greater Glasgow and Clyde Health Board between the years 2001 and 2005. They found that the factors related to lung cancer are smoking and house price, which is used as a proxy measure of deprivation and affluence. In the same paper, they also used a hierarchical Bayesian model to identify boundaries in disease mapping by measuring the similarity and dissimilarity between people living the neighbourhood areas. They believed that the rapid change in disease risk surface mostly occurs when people live in neighbouring areas that very different in social characteristics.

Furthermore, Anderson et al. (2014) aimed to identify clusters of elevated and reduced respiratory disease risk in Glasgow. They proposed a two-stage Bayesian hierarchical agglomerative clustering approach using CAR models. The first stage produced a number of candidate cluster structures for the disease data. The second stage fitted a separate Poisson log linear model to the disease data for each candidate cluster, and the most appropriate cluster structure was selected by minimising the Deviance Information Criterion. This approach allows spatial smoothness within clusters and different levels of average risk between clusters.

However, these approaches still assume that the level of disease risk is constant within each area. Therefore, the novel methodology developed in this thesis tackles this issue by using areal unit data to estimate disease risks at a finer pseudo-continuous spatial scale. This approach also addresses the ecological fallacy where the risk estimates are closer to the individual level and a continuous disease risk map is estimated.

The main aim of this thesis is to make pseudo-continuous spatial inference in disease risk from aggregated areal unit data, and the general approach is outlined as follows. Firstly, regular grid squares are created over the study region, and the areal unit data are transformed to the grid level to estimate disease risks at the grid level. The data needed to do this transformation include the sizes of the intersection areas between each grid square and each areal unit, and the total population in each grid square. Then the expected disease counts and the covariates can be estimated at the grid square level based on the proportion of each grid square that lies in each areal unit. Different types of covariates (e.g. continuous, binary, etc.) are transformed using different approaches. Then the number of disease cases in each grid square is estimated via multinomial sampling. The probabilities of each disease case in an areal unit occurring in a given grid square depends on the size of the intersection area (between areal unit and grid square) and the estimated expected number of disease cases in each grid square based on population size. More details of the methods are discussed in Chapters 4 to 6.

The study region considered in this thesis is the Greater Glasgow and Clyde Health Board, which is in west central Scotland. It is the largest health board in the UK, and provides health care to almost 1.2 million people (https://www.nhsggc.org.uk). Intermediate zones (IZ) are the small area units for which data are available (https://statistics.gov.scot/home), and there are many useful datasets available at this level. There are a total of 257 IZs in the health board, containing populations between 1,074 and 8,989 people with a median population of 4,326. The geographical sizes of these IZs are different, and depends on the density of the population of each IZ. Figure 1.4 shows a map of the IZs in the Greater Glasgow and Clyde Health Board; it can be seen that the geographical areas in the centre of the health board, which are the

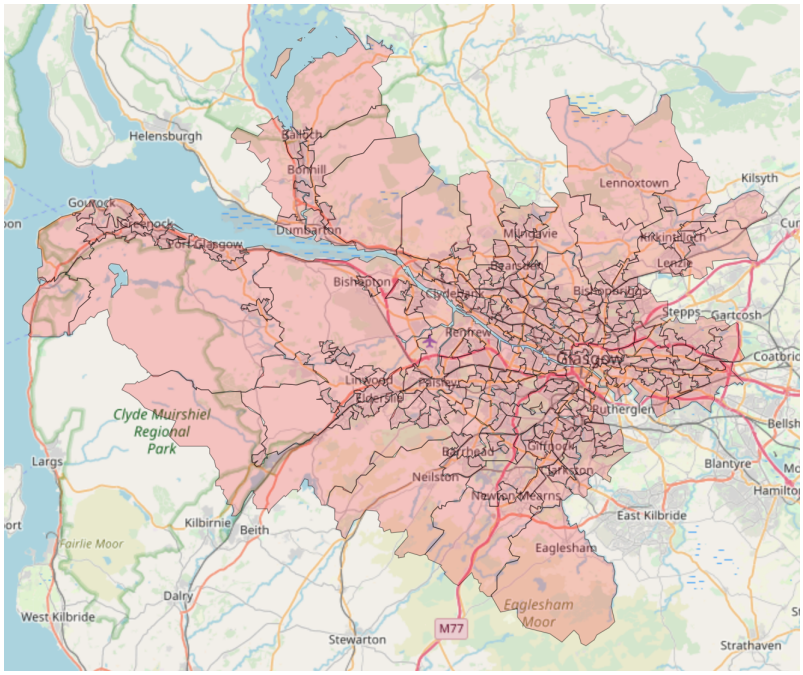


Figure 1.4: Map of the intermediate zones in the Greater Glasgow and Clyde Health Board.

densely populated areas, are much smaller than the more rural areas.

The main focus in this thesis is measuring health inequalities in the Greater Glasgow and Clyde Health Board. This health board area is chosen for a few reasons. First, Figure 1.5 presents male life expectancies in the major cities in the UK, and it can be seen that men in Glasgow have the lowest life expectancy of any city in the UK. Although there is a slightly increased trend for Glasgow, the gaps between Glasgow and the other cities are widening. Moreover, Figure 1.6 illustrates average male life expectancies in the areas close to each of the subway stations in Glasgow. Glasgow Subway has only 15 stations and runs over 10 kilometres with only 24 minutes to complete a circuit. The average life expectancy for men who live in Hillhead (a wealthy area) is 80 years which is higher than those living in Govan (a less wealthy area) by 14 years, even though they live only 3 stations away from each other (6 minutes of travelling). Therefore, it is clear that there are large health inequalities in Glasgow, which should be of interest to study in more detail. The data being used in Chapters 3 to 5 are the total number of hospital admissions for respiratory disease in the 2-year period 2015 - 2016 in each IZ. Additionally, Chapter 6 uses annual counts of disease data for the years 2013 to 2016. Respiratory disease is defined using the International

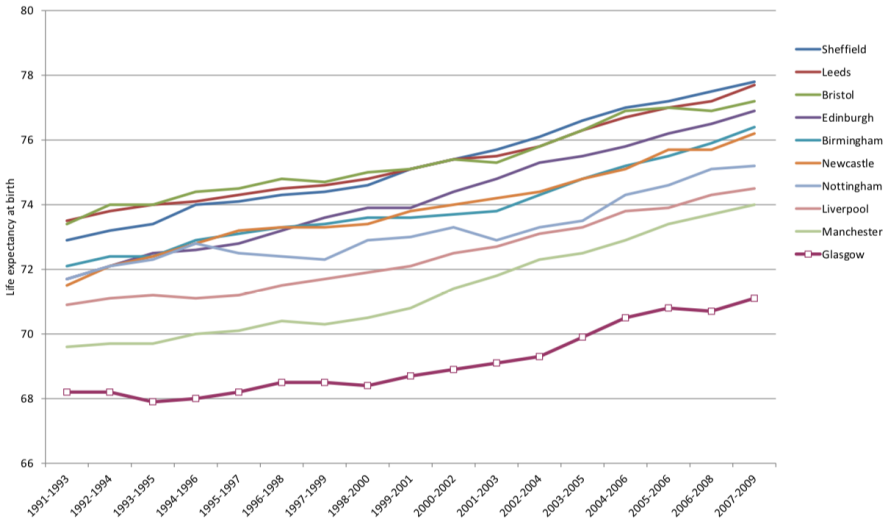


Figure 1.5: Male life expectancy in selected UK cities in the years 1991-93 to 2007-09 (Walsh et al., 2016).

Classification of Diseases volume 10 (ICD10) codes (J00:J99, R09.1). Respiratory disease is selected in this study because the mortality rate in the UK is in the top three highest across Europe for the years 2001 to 2015 (Salciccioli et al., 2018).

The remainder of this thesis is organised as follows. Chapter 2 provides an overview of the statistical methodology which is used in this thesis, including Poisson generalised linear models, Bayesian statistics, spatial statistics as well as the relevant literature in grid square level risk modelling. Specifically in spatial statistics, the most common areal unit models are discussed, including the intrinsic model (Besag et al., 1991), convolution model (Besag et al., 1991), Stern and Cressie model (Stern and Cressie, 2000), and Leroux model (Leroux et al., 2000). Furthermore, spatio-temporal models, which are extended from the spatial models, are outlined including the Bernardinelli model (Bernardinelli et al., 1995), Knorr-Held model (Knorr-Held, 2000), Urgarte model (Ugarte et al., 2012), and Rushworth model (Rushworth et al., 2014). Spatial prediction is also used in this thesis, therefore the method introduced by Krige (1951) is presented in this chapter.

Chapter 3 aims to estimate respiratory disease risk patterns across the Greater Glasgow and Clyde Health Board based on aggregated disease data at the areal unit level. A spatial hierarchical Bayesian model used to achieve this goal is a combination of

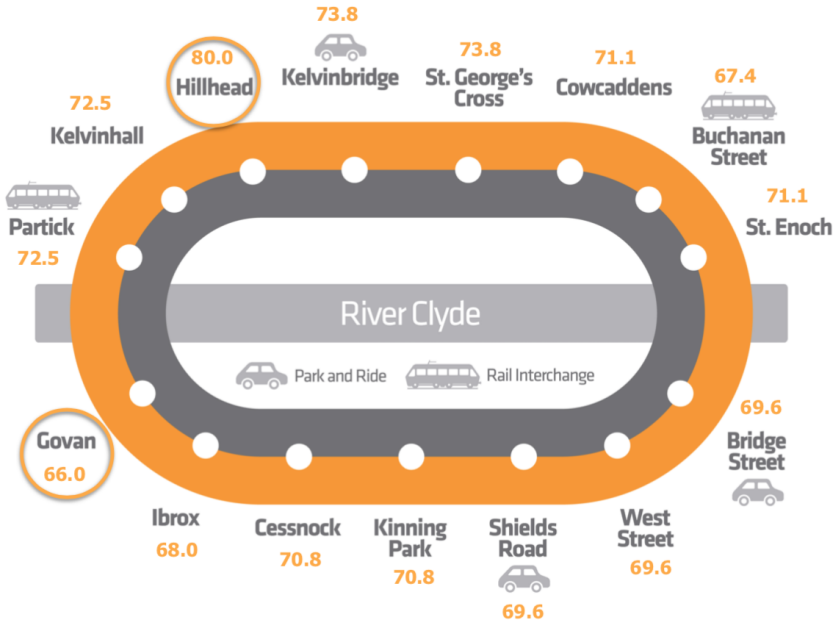


Figure 1.6: Map of subway in Glasgow with male life expectancies in each station (Jack, 2019).

a Poisson generalised linear model and the spatial model proposed by Leroux et al. (2000). This hierarchical model is one of the most commonly used models to estimate the risk at the areal unit level, therefore this chapter gives an overview of the standard methods, limitations, to and how to address these limitations.

Since this thesis focuses on estimating a pseudo-continuous spatial disease risk surface via regular grid squares, Chapter 4 will introduce the methodology of estimating spatial data at the grid level based on areal unit level data via multinomial sampling steps. I then use the multiple imputation approach to estimate pseudo-continuous risk surface utilising the spatial CAR model proposed by Leroux et al. (2000). Unlike the standard approach presented in Chapter 3, this approach allows disease risks to vary within each areal unit or IZ and does not estimate the risks where no people live. A simulation study is conducted to investigate the performance of the models proposed in this chapter, and my chosen model is then applied to the respiratory disease data in the Greater Glasgow and Clyde Health Board.

An alternative approach for estimating a pseudo-continuous risk surface is data augmentation, which is presented in Chapter 5. In this chapter, the CAR model proposed

by Leroux et al. (2000) is utilised to estimate disease risks in each grid square, but the initial results show unrealistic estimates. This is because the Leroux model allows spatial autocorrelation to vary from weak to strong, which could lead to very high variation in the risk estimates. Therefore, the spatial model proposed by Besag et al. (1991) which assumes strong spatial autocorrelation is used instead. Unlike the approach proposed in Chapter 4, this approach estimates disease counts and model parameters via MCMC steps, which allows the uncertainty in the grid level disease cases to be included when estimating model parameters. The performance of the models used in this chapter are examined via a simulation study, then applied to the same dataset as the previous chapter.

Chapter 6 extends the spatial modelling from Chapters 3 to 5 to spatio-temporal modelling, which is used to estimate disease risk at the grid level over time. The main objective of this chapter is to measure health inequalities in the Greater Glasgow and Clyde Health Board and investigate how the disease risk change over time. The spatio-temporal model used in this chapter is proposed by Bernardinelli et al. (1995), which is fitted to respiratory disease data in the Greater Glasgow and Clyde Health Board in the years 2013 to 2016. This model assumes that the trend in disease risks in each area can be explained by a linear relationship, which is suitable for a trend covering four years data. Finally, Chapter 7 summarises the key finding in this thesis and discusses applications for future work.

Chapter 2

Statistical background

2.1 Generalised Linear Model

A generalised linear model (GLM) extends the ordinary regression model for analysing response data $\boldsymbol{y} = (y_1, ..., y_n)$ which are not normally distributed. There are three components; a random component identifies the response variable and its probability distribution, a linear predictor defines the set of covariates, and a link function relates $\mathbb{E}[Y_i]$ to the linear predictor. The random component models independent random variables $\boldsymbol{Y} = (Y_1, ..., Y_n)$ with observations $\boldsymbol{y}$ from the exponential family of distributions

$$\begin{aligned} f(\boldsymbol{Y}) &= f(Y_1, ..., Y_n) \\ &= \prod_{i=1}^{n} f(Y_i), \end{aligned}$$

where

$$f(y_i|\theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]. \tag{2.1.1}$$

Here θ_i is called the *canonical parameter* and represents the location, while ϕ is called the *dispersion parameter* and represents the scale. The linear predictor $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ is given by

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

where $\mathbf{x}_i^\top$ is a vector of covariates and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are the corresponding regression

parameters. The last component of the GLM is the link function that connects the expectation $\mathbb{E}[Y_i]$ to the linear predictor. Let $\mu_i = \mathbb{E}[Y_i]$ for $i = 1, ..., n$. The model links μ_i to η_i by $\eta_i = g(\mu_i)$. Thus, g links μ_i to the covariates through this formula

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

2.1.1 Poisson GLM

In this thesis I will mostly be modelling count data, so therefore the Poisson log-linear model is outlined as it is appropriate distribution for count data. Let the discrete random variable Y_i denote a count of the number of events that occur in an interval of time or space. Then Y_i is a Poisson random variable with sample space $y_i = 0, 1, 2, ...$ and the probability mass function is:

$$f(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \tag{2.1.2}$$

and $\mathbb{E}[Y_i] = \mathrm{Var}[Y_i] = \mu_i$, where μ_i is the mean. The Poisson distribution is a member of the exponential family of distributions because it can be written as (2.1.1) as follows:

$$\begin{aligned} f(Y_i = y_i) &= \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \\ &= \exp[y_i \ln(\mu_i) - \mu_i - \ln(y_i)!]. \end{aligned}$$

Therefore, $\theta_i = \ln(\mu_i)$, $b(\theta_i) = \mu_i = \exp(\theta_i)$, $a(\phi) = 1$ and $c(y_i, \phi) = -\ln(y_i)!$. The response data from the Poisson distribution are non-negative, so the link function that is suitable and commonly used is the log. The basic Poisson GLM can be specified as follows:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i) \qquad i = 1, ..., n, \\ \ln(\mu_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}. \end{aligned} \tag{2.1.3}$$

2.1.2 Maximum likelihood estimation for $\boldsymbol{\beta}$ for the Poisson Model

Suppose that the Poisson random variable Y_i is observed on the ith replicate of an experiment and that Y_i has probability mass function $f(Y_i)$, then assuming independence,

the overall likelihood function of the data is

$$\begin{aligned} f(Y_1 = y_1, ..., Y_n = y_n) &= f(Y_1 = y_1) \times ... \times f(Y_n = y_n) \\ &= f(y_1) \times ... \times f(y_n). \end{aligned}$$

The likelihood function of the unknown parameter $\boldsymbol{\beta}$, given a sample of data, $(y_1, ..., y_n)$ is defined as follows:

$$\begin{aligned} L(\boldsymbol{\beta}; y_1, ..., y_n) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}. \end{aligned}$$

It is often easier to work with the natural logarithm of the likelihood function which is known as the log-likelihood function and is denoted as $l(\boldsymbol{\beta}) = \ln\{L(\boldsymbol{\beta}; \boldsymbol{y})\}$. The log-likelihood of the Poisson distribution is given as follows:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \ln\left[\prod_{i=1}^{n} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}\right] \\ &= \sum_{i=1}^{n} \ln\left[\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}\right] \\ &= \sum_{i=1}^{n} \left[-\mu_i + y_i \ln(\mu_i) - \ln(y_i!)\right] \\ &= \sum_{i=1}^{n} [-\exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \ln(y_i!)]. \end{aligned}$$

To find the MLE for $\boldsymbol{\beta}$, differentiation of the log-likelihood function is needed

$$\begin{aligned} \frac{d(\boldsymbol{\beta})}{d\beta_j} &= \sum_{i=1}^{n} \left[-\mathrm{x}_{ij} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + y_i \mathrm{x}_{ij}\right] \\ &= \sum_{i=1}^{n} \left[\mathrm{x}_{ij} \left(y_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\right)\right]. \end{aligned}$$

To find the turning point set $d(\boldsymbol{\beta})/d\beta_j = 0$ for all $j = 1, ..., p$

$$\begin{pmatrix} \sum_{i=1}^{n} \mathrm{x}_{i1}(y_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \sum_{i=1}^{n} \mathrm{x}_{i2}(y_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \vdots \\ \sum_{i=1}^{n} \mathrm{x}_{ip} y_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{n} \mathrm{x}_{i1} y_i - \sum_{i=1}^{n} \mathrm{x}_{i1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \sum_{i=1}^{n} \mathrm{x}_{i2} y_i - \sum_{i=1}^{n} \mathrm{x}_{i2} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \vdots \\ \sum_{i=1}^{n} \mathrm{x}_{ip} y_i - \sum_{i=1}^{n} \mathrm{x}_{ip} \exp(\mathbf{x_i}^\top \boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{n} \mathrm{x}_{i1} y_i \\ \sum_{i=1}^{n} \mathrm{x}_{i2} y_i \\ \vdots \\ \sum_{i=1}^{n} \mathrm{x}_{ip} y_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \mathrm{x}_{i1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \sum_{i=1}^{n} \mathrm{x}_{i2} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ \vdots \\ \sum_{i=1}^{n} \mathrm{x}_{ip} \exp(\mathbf{x_i}^\top \boldsymbol{\beta}) \end{pmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \exp(\boldsymbol{X}\boldsymbol{\beta})$$

where $$\boldsymbol{X} = \begin{pmatrix} 1 & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1p} \\ 1 & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{n2} & \cdots & \mathrm{x}_{np} \end{pmatrix}.$$

This equation cannot be solved analytically to find the MLE for $\boldsymbol{\beta}$. Therefore, numerical optimization can be used by applying the Newton-Raphson method with Fisher scoring which is equivalent to the iteratively reweighted least square (IRWLS) algorithm.

Set $\boldsymbol{\beta}^{(0)}$ as a starting value, then compute the next estimate $\boldsymbol{\beta}^{(j+1)}$ from $\boldsymbol{\beta}^{(j)}$ as

$$\boldsymbol{\beta}^{(j+1)} = \left[\boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(j)}) \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(j)}) \widetilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(j)}),$$

where

$$\begin{aligned} \boldsymbol{W}(\boldsymbol{\beta}^{(j)}) &= \text{diag}(W_{11}(\boldsymbol{\beta}^{(j)}), ..., W_{nn}(\boldsymbol{\beta}^{(j)}))_{n\times n} \\ W_{ii}(\boldsymbol{\beta}^{(j)}) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(j)}) \\ \widetilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(j)}) &= \boldsymbol{X}\boldsymbol{\beta} + \begin{pmatrix} \frac{y_1 - \exp(\mathbf{x}_1\boldsymbol{\beta})}{\exp(\mathbf{x}_1\boldsymbol{\beta})} \\ \vdots \\ \frac{y_n - \exp(\mathbf{x}_n\boldsymbol{\beta})}{\exp(\mathbf{x}_n\boldsymbol{\beta})} \end{pmatrix}. \end{aligned}$$

The algorithm stops when $\|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\| < \delta$ for some small value δ.

2.2 Bayesian Statistics

Bayesian statistics is a branch of statistics that provides people with the tools to update their beliefs in the evidence of new data. Bayes' theorem was developed by Thomas Bayes in the 18th century (Bayes, 1764) and is defined for events A and B as follows:

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

where $P(A|B)$ is the conditional probability of event A happening given that event B has happened, and $P(A), P(B)$ are the probabilities of events A and B happening. This theorem can be adapted to provide a basis for model parameter inference. In frequentist statistics, the data $\boldsymbol{Y}$ are used to estimate parameters $\boldsymbol{\theta}$, but in Bayesian statistics each parameter can be assigned in advance a distribution which is known as a prior distribution, $f(\boldsymbol{\theta})$. A prior distribution can be used to reflect prior beliefs about a parameter and it can be updated by using the observed data, $\boldsymbol{Y}$, via the data likelihood, $f(\boldsymbol{Y}|\boldsymbol{\theta})$, in order to define a posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{Y})$ as follows

$$f(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{Y})},$$

where $f(\boldsymbol{Y})$ is the marginal distribution of the observed data. However the distribution $f(\boldsymbol{Y})$ is generally difficult to estimate, as it usually involves a multidimensional integral.

As it does not depend on $\boldsymbol{\theta}$, the posterior distribution can instead be written up to proportionality as

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

the product of the data likelihood and the prior distribution.

2.2.1 Prior Distributions

The prior distribution $f(\boldsymbol{\theta})$, often simply called the prior, represents the information about uncertain parameters $\boldsymbol{\theta}$, that are combined with the probability distribution of the observed data $\boldsymbol{Y}$ to determine the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y})$. The prior should be determined before seeing the data, represents prior beliefs, and it could be based on information from a previous study or an expert in the field. It can be chosen in various forms depending on the type of data and model. Furthermore, if the prior and the posterior distribution are from the same family, which means that the prior and the posterior have the same form, the prior is called a conjugate prior. As an example

suppose $Y \sim \text{Poisson}(\theta)$, then the likelihood is:

$$\begin{aligned} L(\theta|y) &= f(y|\theta) \\ &= \frac{\exp(-\theta)\theta^y}{y!} \\ L(\theta|y) &\propto \exp(-\theta)\theta^y. \end{aligned}$$

If I select a Gamma distribution as the prior for θ

$$\begin{aligned} \theta &\sim \text{Gamma}(\alpha, \beta) \\ f(\theta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta), \quad \theta > 0 \\ &\propto \theta^{\alpha-1}\exp(-\beta\theta). \end{aligned}$$

The posterior can therefore be written as

$$\begin{aligned} f(\theta|y) &\propto f(y|\theta)f(\theta) \\ &\propto \exp(-\theta)\theta^y \times \theta^{\alpha-1}\exp(-\beta\theta) \\ &\propto \theta^{\alpha+y-1}\exp[-(\beta+1)\theta]. \end{aligned}$$

Thus the posterior distribution will be of the same general form as the prior distribution, that is a Gamma distribution,

$$f(\theta|y) \sim \text{Gamma}(\alpha+y, \beta+1). \tag{2.2.1}$$

A Gamma prior is the conjugate prior for a Poisson mean parameter resulting in a Gamma distribution for the posterior distribution. Conjugate priors are popular as they allow the posterior to be a known distribution, which makes inference easier. Note that conjugate priors can not be used all the time since some distributions do not have conjugate priors.

The selection of the prior will influence the posterior distribution so it is important to choose a sensible prior. If a prior has no impact on the posterior distribution it is called a noninformative prior. An example of noninformative prior is Jeffreys prior (Jeffreys, 1946), which is designed to be constant under reparameterisation. Jeffreys priors are formed as $f(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}$ where $I(\boldsymbol{\theta})$ is the Fisher information, defined as

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

In contrast, an informative prior is a prior that is not dominated by the likelihood and that has influence on the posterior distribution. For example, a Gaussian distribution with variance one ($\theta_k \sim \text{N}(0,1)$) for real valued parameters. It is also possible to assign a weakly informative prior when we do not have much information about a parameter e.g. $\theta_k \sim \text{N}(0, 1000000)$. Furthermore, if we make α and β from (2.2.1) really small values then they have very little influence on the posterior distribution, and in such a case, the Gamma prior would also be weakly informative. A prior is called an improper

prior when the integral of the prior function is not finite. However, an improper prior can be used, as long as the resulting posterior is proper.

2.2.2 Markov Chain Monte Carlo Simulation

Markov Chain Monte Carlo (MCMC) simulation is the most commonly used approach to estimate parameters in a Bayesian model. It is a simulation based approach that is used when $\boldsymbol{\theta}$ cannot be sampled directly from $f(\boldsymbol{\theta}|\boldsymbol{Y})$. Two main algorithms have been proposed, the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings algorithm (Hastings, 1970).

Gibbs Sampler

The idea in Gibbs sampling is to generate posterior samples by sweeping through each variable (or block of variables) to sample from its conditional distribution with the remaining variables fixed at their current values. The Gibbs sampling algorithm for drawing d samples from the posterior distribution is as follows.

Gibbs sampler algorithm

1. Set initial values $\boldsymbol{\theta}^{(0)}=(\boldsymbol{\theta}_1^{(0)},...,\boldsymbol{\theta}_b^{(0)})$, where $\boldsymbol{\theta}$ has been partitioned in to b sets of parameters $(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_b)$.

2. For each iteration $i = 1,...,d$ generate $\boldsymbol{\theta}_k^{(i)}$ for each $k = 1,...,b$ in turn from the conditional distribution $f(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(i)},...,\boldsymbol{\theta}_{k-1}^{(i)},\boldsymbol{\theta}_{k+1}^{(i-1)},...,\boldsymbol{\theta}_b^{(i-1)},\boldsymbol{Y})$.

Gibbs sampling works when each $f(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(i)},...,\boldsymbol{\theta}_{k-1}^{(i)},\boldsymbol{\theta}_{k+1}^{(i-1)},...,\boldsymbol{\theta}_b^{(i-1)},\boldsymbol{Y})$ is a proper distribution. However, there are some limitations to the Gibbs sampler. If the conditional distribution is not straightforward to sample from, simulation is generally conducted by using a more complex method, such as the Metropolis-Hastings algorithm.

Metropolis-Hastings

The Metropolis-Hastings algorithm generates a sequence of random samples from a posterior distribution for which direct sampling is difficult. The Metropolis-Hastings algorithm for drawing d samples from the posterior distribution is as follows.

Metropolis-Hastings algorithm

1. Set initial values $\boldsymbol{\theta}^{(0)}=(\boldsymbol{\theta}_1^{(0)},...,\boldsymbol{\theta}_b^{(0)})$.

2. For each iteration $i = 1,...,d$ in turn generate $\boldsymbol{\theta}_k^{(i)}$ for each $k = 1,...,b$ sets of parameters using the following steps.

   (a) Generate a set of proposed parameter values $\boldsymbol{\theta}_k^*$ from a proposal distribution $g(\boldsymbol{\theta}_k^*|\boldsymbol{\theta}_k^{(i-1)})$.

   (b) Accept the proposed set of values $\boldsymbol{\theta}_k^*$ with probability

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

   and reject with probability 1-p.

   (c) If we accept the proposal then set $\boldsymbol{\theta}_k^{(i)} = \boldsymbol{\theta}_k^*$, otherwise set $\boldsymbol{\theta}_k^{(i)} = \boldsymbol{\theta}_k^{(i-1)}$.

The selection of starting values $\boldsymbol{\theta}^{(0)}$ may influence the performance of the simulation (Turner et al., 2013). The starting values can take any possible value, but should be chosen carefully to ensure that they are not too large or too small. For example, if the true value of μ is 0 and you set the initial value $\mu^{(0)}$=1,000,000, then it will take a large number of steps to reach the true value. One approach for selecting starting values is to run multiple chains with different initial values, while another is to estimate a reasonable range of starting values from the data. In step 2 (b), in the Metropolis-Hastings algorithm the proposal distribution g needs to be assigned. The most commonly used proposal for Metropolis-Hastings is a Gaussian distribution with mean $\boldsymbol{\theta}_k^{(i-1)}$ (the current value) and variance $\boldsymbol{V}$ because it is symmetric; that is $g(\boldsymbol{\theta}_k^{(i-1)}|\boldsymbol{\theta}_k^*) = g(\boldsymbol{\theta}_k^*|\boldsymbol{\theta}_k^{(i-1)})$. This is a special case of the Metropolis-Hastings algorithm and is called the Metropolis algorithm (Metropolis et al., 1953). Therefore, the acceptance probability is equal to $p = \min\left\{1, \frac{f(\boldsymbol{\theta}_k^*|\boldsymbol{Y})}{f(\boldsymbol{\theta}_k^{(i-1)}|\boldsymbol{Y})}\right\}$.

Proposing small moves to our chain will likely lead to more of our proposals being accepted, which mean we will have a high acceptance rate. In contrast, if we propose moves which are too large, the proposals are often rejected and our acceptance rate will be low. Roberts et al. (1997) suggested that it is optimal to have an acceptance rate

close to 0.25 for parameters of high dimension and approximately 0.5 for parameters of dimension one or two.

Convergence of the MCMC algorithm

The basic idea of the MCMC algorithm is to construct a Markov chain with properties which allow it to converge to the posterior distribution of interest after a number of iterations. Technically, convergence occurs when the simulated Markov chain converges in distribution to the target distribution. There are many approaches to diagnose whether the samples have converged. Geweke et al. (1991) introduced a convergence diagnostic based on standard time series methods. The chain is divided into two windows containing the first 10% and last 50% of the chain. Ideally, if the whole chain is stationary, the means of both windows should be nearly equal. A Z-statistic is calculated as the difference between the two means divided by the asymptotic standard error of the difference, where the variance is defined by a spectral density estimation. As the number of iterations increases, the distribution of the Z-statistic approaches the $N(0, 1)$ distribution if the chain has converged.

Gelman and Rubin (1992) proposed a convergence test based on multiple chains, where each chain starts from different overdispersed initial values to compute an estimation of a posterior distribution. This approach is based on a comparison of the within and between chain variances for each parameter. A potential scale reduction factor ($PSRF$) is computed for each scalar quantity of interest, which estimates the factor by which the scale of the Student t density that approximates the posterior distribution of a scalar parameter might be reduced if the chains are run to infinity. The $PRSF$ is defined as follows:

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

where B is the variance between the means of the m chains, W is the average of the m within-chain variances, and n is the number of iterations. As n becomes larger, the $PSRF$ is approximately 1 if the algorithm converges. Values < 1.1 suggest convergence.

Convergence is also commonly diagnosed by visual assessment of trace plots, where

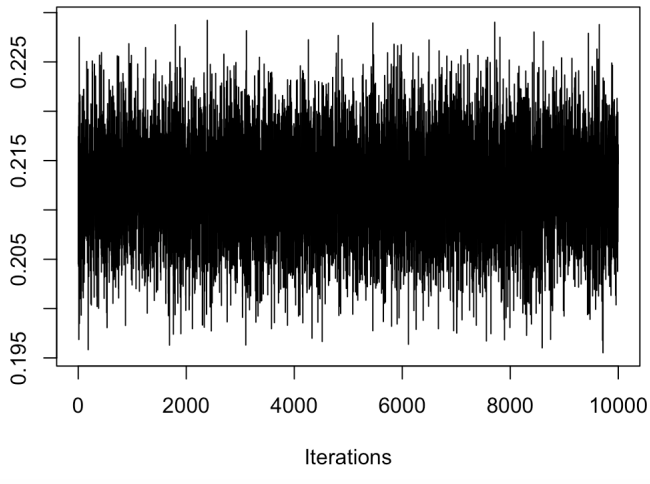


Figure 2.1: An example of a trace plot.

convergence is assumed once the trace plots are considered weakly stationary. A trace plot is a plot representing the generated values of a parameter for each iteration in a chain. Basically, a line connects the points in such a plot for simplicity of considering the path traveled by the chain. An example trace plot for 10,000 MCMC samples is shown in Figure 2.1. In this thesis, convergence will be assessed via trace plots, the Geweke and the Gelman and Rubin approaches.

2.2.3 Inference

Once the Markov chain has converged, subsequent samples come from the posterior distribution. The chain will typically take some time to converge, and we are only interested in the equilibrium distribution, i.e. those samples obtained after convergence. We should therefore remove the samples which were obtained prior to convergence. This is known as the **burn in** period. In addition, samples generated from MCMC algorithms are normally correlated which means that our set of samples are not independent. We can create a set of independent samples by the process of **thinning**, which involves taking only every kth sample from the posterior distribution and discarding all others. The main drawback of thinning is throwing away information. However Link and Eaton (2012) argued that there are a few reasons for thinning, firstly to reduce the autocorrelation between MCMC samples. The effective sample size is the approximate number of independent samples from the MCMC chain, which can be calculated by

$$n_{\text{eff}} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

where n is the number of iterations and ρ_k is the autocorrelation between the samples in the Markov chain at lag k. The second is due to the limitations of computer memory and storage since there are large numbers of MCMC samples to be stored.

A key advantage of Bayesian statistics is the ability to obtain a complete posterior distribution for each parameter. To obtain a point estimate for a univariate θ from its posterior distribution, we can simply apply an appropriate summary statistic such as the posterior mean or median to the MCMC samples. $100(1-\alpha)\%$ posterior uncertainty intervals can also be obtained directly from the MCMC samples, and in Bayesian modelling these are known as credible intervals. To create a $100(1-\alpha)\%$ credible interval from the marginal posterior distribution for the parameter of interest, we simply select the interval between $\alpha/2$ and $1-\alpha/2$ quantiles of the MCMC samples. The interpretation of a credible interval is different from that of a confidence interval in frequentist statistics. A 95% credible interval means that the parameter will lie in the interval with posterior probability 0.95 while a 95% confidence interval means 95% of the intervals produced from repeated sampling of the data will contain the parameter of interest.

2.3 Spatial Statistics

Spatial statistics is the analysis and modelling of data at different locations. Spatial data are any form of statistical data which have geographical locations attached. Spatial data and time series data have the same feature, where observations close in time or space are likely to be similar, while observations further apart are likely to be independent. It is therefore necessary to account for autocorrelation when modelling spatial data. There are three main forms of spatial data; point process data, areal data, and geostatistical data. Point process data are a form of spatial data where the actual location itself is the feature of interest. An example of point process data would be

the locations of trees in a forest. Areal data are based on a geographical area which is divided into non-overlapping areas such as census tracts or electoral wards, and there is one summary measurement per unit. Geostatistical data consist of a set of observations taken at precise spatial locations, such as the concentration of air pollution measured by monitors across Glasgow. In this thesis I am mainly modelling areal unit data in a disease context, so in the remainder of this chapter I discuss the field of areal unit data modelling known as disease mapping.

2.3.1 Disease mapping

Disease mapping involves estimating the spatial pattern in disease risk over a pre-defined study region such as a city or country. The study region is divided in to a number of small non-overlapping areas, $\mathcal{A} = (\mathcal{A}_1, ..., \mathcal{A}_n)$, and we obtain data about the number of people with a particular disease in each area. These data take the form $\boldsymbol{Y} = [Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$, where $Y(\mathcal{A}_i)$ denotes disease counts from people living in area $\mathcal{A}_i$. In general, the areas $\mathcal{A}$ consist of pre-existing administrative units, for example postcode areas, electoral wards or counties. Disease count data for these areas may often be published by a government or health board, for example health data in Scotland are published at www.statistics.gov.scot. Modelling disease risk based on raw disease counts such as these can be misleading because each area varies in terms of population size and demographic structure (e.g. age and sex). For example, as elderly people have a higher risk of heart disease than younger people, then areas which have a higher percentage of elderly people are likely to have a higher number of heart disease cases, but this does not necessarily imply a higher disease risk in such areas. We can address this issue by computing the expected number of disease cases in each area, based on demographic information. The expected disease counts are denoted by $\mathbf{e} = [\mathrm{e}(\mathcal{A}_1), ..., \mathrm{e}(\mathcal{A}_n)]$, and can be constructed via indirect standardisation based on the age and sex of the population within each area. We construct a set of m strata of the population based on sex and age (e.g. female 0-4, female 5-9, etc), and then compute the expected disease counts via $\mathrm{e}(\mathcal{A}_i) = \sum_{j=1}^{m} N_{ij} r_j$, where N_{ij} is the population of area $\mathcal{A}_i$ in strata j, and r_j is the national disease rate in strata j.

We can use these expected disease counts to compute a measure of disease risk known as the standardised incidence ratio (SIR), which is given for area $\mathcal{A}_i$ as $\text{SIR}(\mathcal{A}_i) = \frac{Y(\mathcal{A}_i)}{\text{e}(\mathcal{A}_i)}$. An SIR value equal to 1 means there are as many disease cases as expected in an areal unit. An SIR value greater than 1 means there are more disease cases than expected and hence higher risk, while an SIR value less than 1 means there are fewer disease cases than expected and hence lower risk. For example, an SIR value of 1.2 indicates there are 20% more disease cases than expected, while an SIR value of 0.9 indicates there are 10% fewer disease cases than expected.

2.3.2 The neighbourhood matrix

In order to conduct a spatial analysis for areal unit data, it is necessary to define a neighbourhood matrix, which is a summary of which areas are close to which other areas. In fact the neighbourhood matrix, $\boldsymbol{W}$, defines the spatial autocorrelation structure for the n areas $\{\mathcal{A}_i : i = 1, ..., n\}$. $\boldsymbol{W}$ is typically a symmetric $n \times n$ matrix, with element w_{ij} denoting how close areas $(\mathcal{A}_i, \mathcal{A}_j)$ are. There are different ways to define $\boldsymbol{W}$, and the values of this matrix can be binary or continuous. However $\boldsymbol{W}$ is non-negative, with larger values w_{ij} denoting areas $(\mathcal{A}_i, \mathcal{A}_j)$ being spatially closer than if w_{ij} were smaller. One method of defining a binary specification for $\boldsymbol{W}$ is to set $w_{ij} = 1$ if areas $(\mathcal{A}_i, \mathcal{A}_j)$ are defined to be neighbours and $w_{ij} = 0$ otherwise. The three most common ways to specify neighbours are: (a) if areas $(\mathcal{A}_i, \mathcal{A}_j)$ share a common border; (b) if the centroids of areas $(\mathcal{A}_i, \mathcal{A}_j)$ are within a fixed distance d of each other; and (c) if area $\mathcal{A}_i$ is one of the k closest areas to area $\mathcal{A}_j$ in term of distance. Continuous $\boldsymbol{W}$ matrices are typically based on distance, with an example being the inverse distance between the two areas' centroids; $w_{ij} = 1/d_{ij}$, where d_{ij} is the distance between centroids of areas $(\mathcal{A}_i, \mathcal{A}_j)$, and $w_{ii} = 0$. However, using non-sparse $\boldsymbol{W}$ matrices can be computationally intensive due to the increased complexity involved in fitting the models described in section 2.3.4. In this thesis, we focus on scenarios where $\boldsymbol{W}$ is defined by the binary specifications (a) based on sharing a common border or (c) based on nearest distance.

The k-nearest neighbour weights matrix for areas $(\mathcal{A}_i, \mathcal{A}_j)$ ensures that every region has at least k neighbours, which avoids the issues surrounding isolated regions. However,

a potential problem with k-nearest neighbour weights is the occurrence of ties, i.e., when more than one area has equal distance from area $\mathcal{A}_i$. In such cases, the simplest approach is to randomly choose a nearest neighbour when ties occur. Moreover, a k-nearest neighbour matrix is not necessarily symmetric, area $\mathcal{A}_i$ being one of the k-nearest neighbours to area $\mathcal{A}_j$ does not imply that area $\mathcal{A}_j$ is also one of the k-nearest neighbours to area $\mathcal{A}_i$. There may be another area that is closer to area $\mathcal{A}_j$ than area $\mathcal{A}_i$. This can be resolved by defining $w_{ij} = 1$ if area $\mathcal{A}_i$ is the k-nearest neighbour to area $\mathcal{A}_j$ or area $\mathcal{A}_j$ is the k-nearest neighbour to area $\mathcal{A}_i$ and $w_{ij} = 0$ otherwise

2.3.3 Moran's I test

Consider data $\boldsymbol{Y} = (Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n))$ relating to $(\mathcal{A}_1, ..., \mathcal{A}_n)$ that have been collected (one observation per unit). Then the level of spatial autocorrelation in $\boldsymbol{Y}$ can be calculated via Moran's I statistic (Moran, 1950), which is given as follows:

$$I = \frac{n}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}[Y(\mathcal{A}_i) - \bar{Y}(\mathcal{A})][Y(\mathcal{A}_j) - \bar{Y}(\mathcal{A})]}{\sum_{i=1}^{n}[Y(\mathcal{A}_i) - \bar{Y}(\mathcal{A})]^2}. \tag{2.3.1}$$

The value of Moran's I ranges between -1 and 1. Values close to -1 indicate a strong negative spatial autocorrelation, an example is a checkerboard pattern. Values close to 1 indicate a strong positive spatial autocorrelation, this means values cluster together as similar values are close to each other. Values close to 0 indicate spatial independence or no autocorrelation. However, Moran's I is mostly non-negative since negative spatial autocorrelation is rare in practice. A permutation approach can be used to test the hypothesis of spatial autocorrelation. The hypotheses for the test are

$$H_0 : \text{There is no spatial association}$$
$$\text{and} \quad H_1 : \text{There is some spatial association.}$$

The p-value can be computed by randomly permuting the data set M times, and then calculating the Moran's I value from (2.3.1) for each permuted data set. The p-value is computed as the proportion of those simulated Moran's I values that are more extreme than the observed Moran's I value from the actual data set.

2.3.4 Areal unit modelling

Typically in this thesis the responses are count data, and thus areal data are generally modelled by the Poisson log linear model (2.1.3), which is extended to account for the spatial autocorrelation in the data. The spatial pattern in the data is modelled by a combination of the covariates and a set of random effects. Let $\boldsymbol{X} = (\mathbf{x}_1^\top, ..., \mathbf{x}_n^\top)^\top$ denote the $n \times p$ matrix of covariates including a column of ones for the intercept term, where the values relating to row i for areal unit $\mathcal{A}_i$ are denoted by $\mathbf{x}_i = (1, \mathrm{x}_{i2}, ..., \mathrm{x}_{ip})$, while the vector of random effect terms is denoted by $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$. A Poisson log-linear spatial model is typically used to model count data $\boldsymbol{Y}$, and a general specification is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(e_i\theta_i) \qquad i = 1, ..., n, \\ \ln(\theta_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i, \end{aligned} \tag{2.3.2}$$

where e_i is the expected number of disease cases in area i which can be computed via indirect standardisation and θ_i denotes the disease risk in area i. The random effects $\boldsymbol{\phi}$ are commonly modelled by the class of prior distributions known as conditional autoregressive (CAR) models. These models can be specified by a set of n univariate full conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. The spatial autocorrelation between these random effects is controlled by the neighbourhood matrix $\boldsymbol{W}$, which was specified in Section 2.3.2, with $w_{ij} = 1$ if area $(\mathcal{A}_i, \mathcal{A}_j)$ share a common border and $w_{ij} = 0$ otherwise. A number of conditional autoregressive prior models have been proposed, and the four models that are most commonly used are described below.

Intrinsic Model

The simplest CAR prior is the intrinsic autoregressive model proposed by Besag et al. (1991), and the full conditional distribution is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \; N\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^2}{\sum_{j=1}^n w_{ij}}\right) \qquad i = 1, ..., n. \tag{2.3.3}$$

The conditional expectation of ϕ_i is the mean of the random effects in neigbouring areal

units (as specified by $\boldsymbol{W}$), while the conditional variance is inversely proportional to the number of neighbouring units. The variance formula is appropriate for strong spatial autocorrelation, because the more neighbouring areas there are, the more information there is to estimate the value of the random effect, hence the variance decreases. One disadvantage of this model is that there is no parameter to control the level of the spatial autocorrelation between the random effects: if one multiplies each ϕ_i by 10, τ^2 will increase but the spatial autocorrelation structure does not change. Therefore, the intrinsic model is sensible in cases where the data have strong spatial autocorrelation but is not appropriate for weak and moderate spatial autocorrelation. The joint distribution for $\boldsymbol{\phi}$ corresponding to (2.3.3) is given by

$$\boldsymbol{\phi} \sim N(\mathbf{0}, \tau^2[\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}]^-),$$

where "$-$" denotes a generalised inverse and $\boldsymbol{W}\mathbf{1}$ is the row sum of $\boldsymbol{W}$.

Convolution Model

The convolution model was also proposed by Besag et al. (1991), and is also known as the Besag-York-Mollié (BYM) model. It combines the intrinsic CAR model (2.3.3) with a set of independent random effects. The model from (2.3.2) is extended to be given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(e_i\theta_i) \qquad i = 1, ..., n, \\ \ln(\theta_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i + \psi_i, \\ \psi_i &\sim \text{N}(0, \sigma^2), \end{aligned} \tag{2.3.4}$$

where $\boldsymbol{\phi}$ is a set of random effects from the intrinsic CAR model (2.3.3) and $\boldsymbol{\psi} = (\psi_1, ..., \psi_n)$ is an additional set of independent random effects $\psi_i \sim \text{N}(0, \sigma^2)$. Different levels of spatial autocorrelation can be modelled by varying the relative levels of variation in $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. The major drawback of this model is it is difficult to estimate both random effect sets separately, normally only their sum $\phi_i + \psi_i$ is identifiable.

Stern and Cressie Model

The model proposed by Stern and Cressie (2000) was adapted from the intrinsic CAR model (2.3.3), by adding a spatial autocorrelation parameter ρ into the model. The model is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \; N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right). \tag{2.3.5}$$

The conditional expectation is equal to a proportion of the mean of the random effects in neighbouring units, while the conditional variance is the same as the intrinsic model. The parameter ρ controls the level of the spatial autocorrelation between the random effects, $\rho = 0$ corresponds to independence, while increasing ρ toward one corresponds to increasingly strong spatial autocorrelation. The main disadvantage for this model is the structure of the conditional variance, because in the case where $\rho = 0$ we are assuming independence, but the conditional variance still depends on the number of neighbours.

Leroux Model

The Leroux model was proposed by Leroux et al. (2000). The model is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \; N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}\right). \tag{2.3.6}$$

The conditional expectation is again a function of the random effects in neighbouring areas, while the conditional variance addresses the issues identified in the Cressie model (2.3.5). For example, a value of $\rho = 1$ indicates a strong spatial autocorrelation and corresponds to the intrinsic model (2.3.3), as does (2.3.5). In contrast, if $\rho = 0$ then the random effects are independent and the conditional mean and variance are equal to 0 and τ^2 respectively, so that the variance now does not depend on $\boldsymbol{W}$. The joint distribution for $\boldsymbol{\phi}$ corresponding to (2.3.6) is given by

$$\boldsymbol{\phi} \sim N(\mathbf{0}, \tau^2[\rho(\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}]^{-1}),$$

and the precision matrix $\boldsymbol{Q} = [\rho(\text{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}]$ is invertible if $\rho \in [0, 1)$.

2.3.5 Geostatistical data

In this thesis, I will use geostatistical modelling to predict the data at new locations because I will change the areal unit level data to the grid square level data since I do not have real data at grid square level. This section will introduce theory about geostatistical data. A geostatistical process is a stochastic process

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\},$$

where $Z(\boldsymbol{s})$ is the random variable at location $\boldsymbol{s}$, D is a fixed subset of the p-dimensional space $\mathbb{R}^p$. In this study, p is fixed at $p = 2$ because I have data points with their coordinates (northing and easting), so $D \subset \mathbb{R}^2$. The location $\boldsymbol{s}$ varies continuously across D. However, in practice data are collected from n locations, and are represented by a random variable denoted by $\mathbf{Z} = \{Z(s_1), ..., Z(s_n)\}$. An example of geostatistical data is the concentration of air pollution recorded at monitoring stations in Glasgow. Air pollution exists across the city, but we only receive data at a finite set of locations. The major concerns when modelling geostatistical data are dependence and autocorrelation because non-spatial data typically assume independence of observations. Generally, geostatistical data have positive autocorrelation; the closer in space the locations of two data points are, the more similar their values are likely to be. On the other hand, two locations which are far apart are likely to have less in common. Therefore, the autocorrelation is controlled by the distance between two data locations, and the exponential covariance model will be used in this thesis to capture the autocovariance. In this thesis, the main use of geostatistical processes are for prediction at unobserved locations. We wish to estimate disease risk across the region based on finite observations. In the following section, the theory behind geostatistical processes is introduced.

Mean and autocovariance

The mean and autocovariance functions of the geostatistical process $Z(\boldsymbol{s})$ are given by

$$\begin{aligned}
\mu_Z(\boldsymbol{s}) &= \mathbb{E}[Z(\boldsymbol{s})] \qquad \forall_{\boldsymbol{s}} \in D, \\
C_Z(\boldsymbol{s},\boldsymbol{t}) &= \mathrm{Cov}[Z(\boldsymbol{s}), Z(\boldsymbol{t})] \\
&= \mathbb{E}[(Z(\boldsymbol{s}) - \mu_Z(\boldsymbol{s}))(Z(\boldsymbol{t}) - \mu_Z(\boldsymbol{t}))] \\
&= \mathbb{E}[Z(\boldsymbol{s})Z(\boldsymbol{t})] - \mu_Z(\boldsymbol{s})\mu_Z(\boldsymbol{t}),
\end{aligned}$$

where $\mu_Z(\boldsymbol{s})$ is the mean of the geostatistical process $Z(\boldsymbol{s})$ at location $\boldsymbol{s}$. Here $C_Z(\boldsymbol{s},\boldsymbol{t})$ represents the autocovariance between the data at locations $\boldsymbol{s}$ and $\boldsymbol{t}$. The autocovariance measures the strength of the linear dependence and the directional relationship between $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$. Note that the autocovariance function is symmetric in its arguments, hence $C_Z(\boldsymbol{s},\boldsymbol{t}) = C_Z(\boldsymbol{t},\boldsymbol{s})$ for each $\boldsymbol{s},\boldsymbol{t} \in D$.

Autocorrelation

The autocorrelation measures the strength of the linear association between data at location $\boldsymbol{s}$ and $\boldsymbol{t}$, scaled to be between [-1, 1], and is denoted by $\rho_Z(\boldsymbol{s},\boldsymbol{t})$. The autocorrelation function of the geostatistical process $Z(\boldsymbol{s})$ is given by

$$\begin{aligned}
\rho_Z(\boldsymbol{s},\boldsymbol{t}) &= \mathrm{Corr}[Z(\boldsymbol{s}), Z(\boldsymbol{t})] \\
&= \frac{C_Z(\boldsymbol{s},\boldsymbol{t})}{\sqrt{C_Z(\boldsymbol{s},\boldsymbol{s})C_Z(\boldsymbol{t},\boldsymbol{t})}}.
\end{aligned}$$

Weakly stationary and isotropy

The geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is stationary when if the set of locations in space D are moved by a specific amount in a specified direction, the entire process retains the same characteristics e.g. a constant mean, a constant variance. It does not mean that the data are all the same. A geostatistical process $Z(\boldsymbol{s})$ is strictly stationary if

$$f(Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n)) = f(Z(\boldsymbol{s}_1 + \boldsymbol{h}), ..., Z(\boldsymbol{s}_n + \boldsymbol{h})),$$

for any displacement vector $\boldsymbol{h} \in \mathbb{R}^2$. However, this assumption is often too strict and hard to verify, therefore one often tests whether it is weakly stationary which can be

defined as follows. A geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is weakly stationary if

1. $\mathbb{E}[Z(\boldsymbol{s})] = \mu_Z(\boldsymbol{s}) = \mu_Z$ for some constant mean μ_Z which does not depend on $\boldsymbol{s}$.
2. $\mathrm{Cov}\,(Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h})) = C_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = C_Z(\boldsymbol{h})$, a finite constant that can depend on $\boldsymbol{h}$ but not on $\boldsymbol{s}$.

In addition, if the geostatistical process is invariant to the direction of the displacement $\boldsymbol{h}$ in that only the magnitude of the displacement matters, then the process is isotropic and, the autocovariance function $C_Z(\boldsymbol{h})$ can be simplified to

$$C_Z(\boldsymbol{h}) = C_Z(\|\boldsymbol{h}\|), \tag{2.3.7}$$

where $h = \|\boldsymbol{h}\|$ denotes the Euclidean distance between two locations $\boldsymbol{h} = (h_1, h_2)$ in the geostatistical process, which can be computed by $\|\boldsymbol{h}\| = \sqrt{h_1^2 + h_2^2}$.

Semi-variogram

In geostatistics, the semi-variogram is often used to quantify the autocorrelation in the data. The semi-variogram function of the geostatistical process $Z(\boldsymbol{s})$ is defined as

$$\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{2}\mathrm{Var}[Z(\boldsymbol{s}) - Z(\boldsymbol{t})],$$

which measures the variance of the data difference at two spatial locations $\boldsymbol{s}$ and $\boldsymbol{t}$. Note that $2\gamma_Z(\boldsymbol{s}, \boldsymbol{t})$ is called the variogram. The relationship between the autocovariance and the semi-variogram is explained as follow:

$$\begin{aligned}
\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) &= \frac{1}{2}\mathrm{Var}[Z(\boldsymbol{s}) - Z(\boldsymbol{t})] \\
&= \frac{1}{2}[\mathrm{Var}(Z(\boldsymbol{s})) + \mathrm{Var}(Z(\boldsymbol{t})) - 2\mathrm{Cov}(Z(\boldsymbol{s}), Z(\boldsymbol{t}))] \\
&= \frac{1}{2}[\mathrm{Cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s})) + \mathrm{Cov}(Z(\boldsymbol{t}), Z(\boldsymbol{t})) - 2\mathrm{Cov}(Z(\boldsymbol{s}), Z(\boldsymbol{t}))] \\
&= \frac{1}{2}[C_Z(\boldsymbol{s}, \boldsymbol{s}) + C_Z(\boldsymbol{t}, \boldsymbol{t}) - 2C_Z(\boldsymbol{s}, \boldsymbol{t})].
\end{aligned}$$

Let us denote $\boldsymbol{t} = \boldsymbol{s} + \boldsymbol{h}$, then $\boldsymbol{h}$ is called displacement or spatial lag. Furthermore we assume that geostatistical process $Z(\boldsymbol{s}) \in D$ is weakly stationary. Hence $\mathbb{E}[Z(\boldsymbol{s})] = \mu_Z$

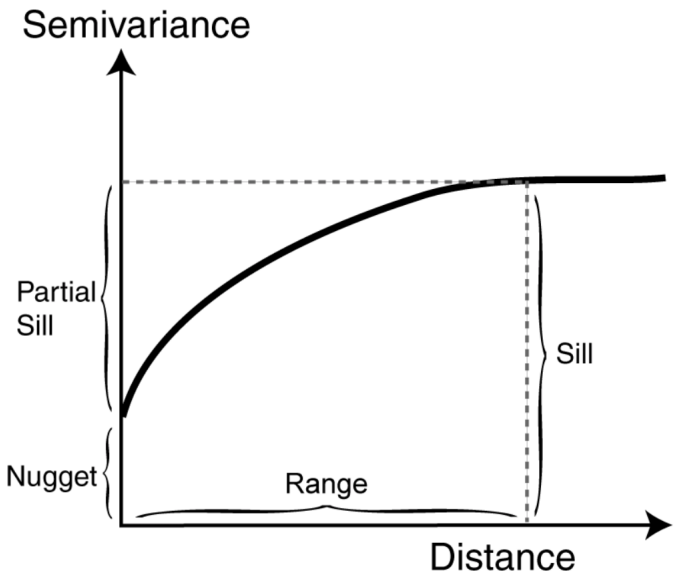


Figure 2.2: The general shape of a semi-variogram.

and $C_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = C_Z(\boldsymbol{h})$, thus we can simplify the semi-variogram to

$$\begin{aligned}
\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) &= \frac{1}{2}[C_Z(\boldsymbol{s}, \boldsymbol{s}) + C_Z(\boldsymbol{s}+\boldsymbol{h}, \boldsymbol{s}+\boldsymbol{h}) - 2C_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h})] \\
&= \frac{1}{2}[C_Z(\boldsymbol{0}) + C_Z(\boldsymbol{0}) - 2C_Z(\boldsymbol{h})] \\
&= C_Z(\boldsymbol{0}) - C_Z(\boldsymbol{h}) \\
&= \gamma_Z(\boldsymbol{h}).
\end{aligned}$$

If the process is also isotropic then $\gamma_Z(\boldsymbol{h}) = \gamma_Z(h = \|\boldsymbol{h}\|)$, where $h = \|\boldsymbol{h}\|$ denotes the Euclidean distance between two locations. Therefore, the semi-variogram can be calculated given the autocovariance. The semi-variogram has the general shape shown in Figure 2.2 under weak stationarity and isotropy. This semivariogram chart is taken from Scheeres, Annaka (2016). A semi-variogram is based on three parameters

1. Partial sill (σ^2) measures the amount of spatially correlated variation in the data.
2. Nugget (ν^2) measures the amount of non-spatial variation or random error.
3. Range (δ) controls the smallest distance (h) at which data become uncorrelated.

There are several weakly stationary and isotropic parametric models that can be used to model geostatistical data, for example exponential, Gaussian, and spherical. The

most commonly used is the exponential autocovariance function, which is given as follows:

$$C_Z(h) = \begin{cases} \sigma^2 \exp(-h/\delta) & h > 0 \\ \nu^2 + \sigma^2 & h = 0, \end{cases}$$

and the associated semi-variogram is

$$\gamma_Z(h) = \begin{cases} \nu^2 + \sigma^2 \exp(-h/\delta) & h > 0 \\ 0 & h = 0. \end{cases}$$

Geostatistical modelling

Suppose we have spatial response data $\mathbf{Z} = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))$ and covariate risk factors $\mathbf{x}(\boldsymbol{s}_i) = (\mathrm{x}_1(\boldsymbol{s}_i), ..., \mathrm{x}_p(\boldsymbol{s}_i))$ for $i = 1, ..., n$, where $\mathrm{x}_1(\boldsymbol{s}_i) = 1$ is the intercept term. All covariates in all n locations are contained in the matrix $\boldsymbol{X}_{n\times p}$. Thus the Gaussian geostatistical model is written as

$$\mathbf{Z} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2.3.8}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is a vector of regression parameters which we want to estimate and forms a linear regression model. $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix related to the autocovariance function $C_Z(h)$ and the vector $\boldsymbol{\theta} = (\sigma^2, \nu^2, \delta)$ includes partial sill, nugget, and range parameters. In this study, a constant mean model is used as the special case which can be written as

$$\mathbf{Z} \sim \mathrm{N}(\beta_0 \mathbf{1}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2.3.9}$$

where $\mathbf{1}_{n\times 1} = (1, ..., 1)$ while $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ii} = \mathrm{Var}(Z(\boldsymbol{s}_i))$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = \mathrm{Cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j))$. We can obtain the parameters estimates $(\hat{\beta}_0, \hat{\boldsymbol{\theta}})$ by maximum likelihood estimation.

Parameter estimation

Consider geostatistical data $\mathbf{Z} = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))$ which has the general form as follows;

$$\mathbf{Z} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2.3.10}$$

and there are two sets of parameters needing to be estimated; parameters in the mean model $\boldsymbol{\beta}$ and parameters in the covariance model $\boldsymbol{\theta} = (\sigma^2, \nu^2, \delta)$. Here these parameters are estimated via a maximum likelihood approach. Assuming for exposition that an exponential autocovariance model is used and the model is given by

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2 \exp(\mathbf{D}/\delta) + \nu^2 \mathbf{I},$$

where $\mathbf{D} = (d_{ij})$ is an $n \times n$ distance matrix with values $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ and I is the $n \times n$ identity matrix. The likelihood function for this multivariate normal distribution is as follows:

$$f(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathrm{Z}) = (2\pi)^{-\frac{n}{2}} \left|\sigma^2 \exp(-\mathbf{D}/\delta) + \nu^2 \mathbf{I}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^\top [\sigma^2 \exp(-\mathbf{D}/\delta) + \nu^2 \mathbf{I}]^{-1} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Then removing unnecessary constants the log-likelihood function is given by

$$\ln[f(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{Z})] = -\frac{1}{2} \ln\left|\sigma^2 \exp(-\mathbf{D}/\delta) + \nu^2 \mathbf{I}\right| - \frac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^\top [\sigma^2 \exp(-\mathbf{D}/\delta) + \nu^2 \mathbf{I}]^{-1} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}).$$

To make the estimation easier the transformation $\xi^2 = \nu^2/\sigma^2$ is applied. Then replacing ν^2 by $\xi^2\sigma^2$ the log-likelihood is given by

$$\ln[f(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{Z})] = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \ln\left|\exp(-\mathbf{D}/\delta) + \xi^2 \mathbf{I}\right| - \frac{1}{2\sigma^2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^\top [\exp(-\mathbf{D}/\delta) + \xi^2 \mathbf{I}]^{-1} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.3.11}$$

Then differentiate the log-likelihood function above with respect to β_0 and σ^2 and then setting equal to zero and solving gives the estimates of $\boldsymbol{\beta}$ and σ^2 as follows;

$$\hat{\boldsymbol{\beta}}(\delta, \xi^2) = (\boldsymbol{X}^\top \mathbf{V}(\delta, \xi^2)^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{V}(\delta, \xi^2)^{-1} \mathbf{Z},$$
$$\hat{\sigma}^2(\boldsymbol{\beta}, \delta, \xi^2) = \frac{1}{n-p} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^\top \mathbf{V}(\delta, \xi^2)^{-1} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}),$$

where p is the number of parameters in the mean model which is subtracted from n in $\hat{\sigma}^2(\boldsymbol{\beta}, \delta, \xi^2)$ to ensure that the estimate is unbiased. Note that I denote $\mathbf{V}(\delta, \xi^2) = \exp(-\mathbf{D}/\delta) + \xi^2 \mathbf{I}$ for simplicity of notation. For the estimation of (ϕ, ξ^2), differentiation of the log-likelihood functions does not work since both parameters are included in $\mathbf{V}(\delta, \xi^2)$, which is then inverted. Hence no closed form solution exists. Therefore, the

estimates $\hat{\boldsymbol{\beta}}(\delta, \xi^2)$ and $\hat{\sigma}^2(\delta, \xi^2)$ are plugged into the log-likelihood function (2.3.11), obtaining the following reduced form (only components that are based on (δ, ξ^2) are included) as follows;

$$\ln[f(\delta, \xi^2)|\mathbf{Z}] = -\frac{n}{2}(\hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \delta, \xi^2)) - \frac{1}{2}\ln(|\mathbf{V}(\delta, \xi^2)|).$$

The estimates $(\hat{\delta}, \hat{\xi}^2)$ can be obtained by numerical maximisation methods using a computer. Then the final estimates of $\hat{\boldsymbol{\beta}}(\delta, \xi^2)$ and $\hat{\sigma}^2(\delta, \xi^2)$ are given by

$$\hat{\boldsymbol{\beta}}(\hat{\delta}, \hat{\xi}^2) = (\boldsymbol{X}^\top \mathbf{V}(\hat{\delta}, \hat{\xi}^2)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbf{V}(\hat{\delta}, \hat{\xi}^2)^{-1}\mathbf{Z},$$
$$\hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \hat{\delta}, \hat{\xi}^2) = \frac{1}{n-H}(\mathbf{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \mathbf{V}(\hat{\delta}, \hat{\xi}^2)^{-1}(\mathbf{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

Spatial prediction

Spatial prediction is the process of predicting the geostatistical processes $Z(s_0)$ at an unmeasured location s_0. There are several methods that can be used for spatial prediction, one of the most common approaches is Kriging, named after Krige (1951) who worked in the mining industry in South Africa. Kriging assumes that the distance between two locations in the study area reflects the spatial autocorrelation between the values of the geostatistical process. It is based on the following result.

Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ be a vector of length n that is split into sub-vectors of length $(q, n-q)$. Then assume that $\boldsymbol{X}$ is multivariate Gaussian distributed, so we have that

$$\boldsymbol{X} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Then the conditional distribution of $\boldsymbol{X}_1|\boldsymbol{X}_2$ is

$$\boldsymbol{X}_1|\boldsymbol{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \tag{2.3.12}$$

Now suppose we have a constant mean for our data, and we propose the geostatistical model

$$\mathbf{Z} \sim (\beta_0 \mathbf{1}, \boldsymbol{\Sigma}(\boldsymbol{\theta})). \tag{2.3.13}$$

Then consider the joint geostatistical process at the n data locations and N prediction locations $\boldsymbol{s}^* = (s_1^*, ..., s_N^*)$:

$$\mathbf{Z}^* = \begin{bmatrix} \mathbf{Z}(\boldsymbol{s}^*) \\ \mathbf{Z} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta_0 \mathbf{1} \\ \beta_0 \mathbf{1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^*(\boldsymbol{\theta}) & \boldsymbol{C}(\boldsymbol{s}^*, \boldsymbol{\theta})^\top \\ \boldsymbol{C}(\boldsymbol{s}^*, \boldsymbol{\theta}) & \boldsymbol{\Sigma}(\boldsymbol{\theta}) \end{bmatrix} \right),$$

where $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}) = \mathrm{Var}(\mathbf{Z}(\boldsymbol{s}^*)) = \sigma^2 \boldsymbol{I} + \nu^2 \boldsymbol{I}$ based on the exponential autocovariance and $\boldsymbol{C}(\boldsymbol{s}^*, \boldsymbol{\theta}) = (\mathrm{Cov}(Z(\boldsymbol{s}_1), Z(\boldsymbol{s}^*), ..., \mathrm{Cov}(Z(\boldsymbol{s}_n), Z(\boldsymbol{s}^*))$. Then using the result (2.3.12) above gives

$$\mathbf{Z}(\boldsymbol{s}^*)|\mathbf{Z} \sim \mathrm{N}\left(\mathbb{E}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}], \mathrm{Var}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}] \right),$$

where

$$\mathbb{E}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}] = \hat{\beta}_0 \mathbf{1} + \boldsymbol{C}(\boldsymbol{s}^*, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1} (\mathbf{Z} - \hat{\beta}_0 \mathbf{1}) \text{ and}$$
$$\mathrm{Var}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}] = \boldsymbol{\Sigma}^*(\hat{\boldsymbol{\theta}}) - \boldsymbol{C}(\boldsymbol{s}^*, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{C}(\boldsymbol{s}^*, \hat{\boldsymbol{\theta}}).$$

Here $(\hat{\beta}_0, \hat{\boldsymbol{\theta}})$ have been estimated by maximum likelihood estimation, and the ordinary kriging predictor is given by

$$\widehat{\mathbf{Z}(\boldsymbol{s}^*)} = \hat{\beta}_0 \mathbf{1} + \boldsymbol{C}(\boldsymbol{s}^*, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1} (\mathbf{Z} - \hat{\beta}_0 \mathbf{1}), \tag{2.3.14}$$

and a corresponding 95% prediction interval is

$$\mathbb{E}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}] \pm 1.96 \sqrt{\mathrm{Var}[\widehat{\mathbf{Z}(\boldsymbol{s}^*)}|\mathbf{Z}]}. \tag{2.3.15}$$

2.4 Spatio-temporal modelling

The spatial modelling approaches outlined in Section 2.3.4 are used to fit data at a single time point in order to investigate disease risk patterns over the study region. However, in some cases, disease data are collected across multiple time points and therefore

spatio-temporal modelling approaches have been developed in order to identify changes in disease risk in both space and time across the area of interest. One of the main aims of these spatio-temporal approaches is to estimate the variation or trends in disease risk over time in different areal units. In this section, I introduce some of the important spatio-temporal models, which are used to achieve different goals of the studies. Since my thesis focuses on count data, all models are written in terms of a Poisson likelihood.

2.4.1 Bernardinelli model

Bernardinelli et al. (1995) proposed a Poisson GLM with the linear predictor including separate parameters for space and time effects and also interactions between space and time. Suppose the response data for area i take the form $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iT})$ for each area $i = 1, ..., n$ for time $t = 1, ..., T$. The model is given by

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(e_{it}\theta_{it}) \qquad i = 1, ...n, \qquad t = 1, ..., T \\ \log(\theta_{it}) &= (\mu + \phi_i) + (\beta + \delta_i)t, \end{aligned} \tag{2.4.1}$$

where Y_{it} is the number of disease cases in area i during time t, which is assumed to follow a Poisson distribution with expected number of disease cases e_{it} and disease risk θ_{it}. μ is a global intercept which is common for all areas and β is an overall slope parameter. This model allows different areas to have different intercepts and slopes by containing random effect terms ϕ_i and δ_i which respectively represent area specific intercepts and linear slopes for area i. In other words, the intercept for area i can be computed by $\mu + \phi_i$ and the slope for area i is $\beta + \delta_i$. Additionally, the random effect terms $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$ are modelled via the conditional autoregressive models outlined in Section 2.3.4, thus allowing for spatial autocorrelation in the area specific intercepts and slopes. The major drawback for this model is it assumes the temporal trend is linear, therefore it does not allow more flexible trends to be estimated.

2.4.2 Knorr-Held model

Knorr-Held (2000) proposed a hierarchical Bayesian model for representing the space-time variation in disease risk. The model contains separate spatial and temporal effects

as well as a space-time interaction term. A Poisson likelihood equivalent of the Knorr-Held model is given by

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(e_{it}\theta_{it}) \qquad i = 1, ...n, \qquad t = 1, ..., T \\ \log(\theta_{it}) &= \mu + \alpha_i + \phi_i + \beta_t + \delta_t + \eta_{it}, \end{aligned} \tag{2.4.2}$$

where μ is the global intercept, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$ and $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are spatial effect terms which account for the spatial structure of the data, $\boldsymbol{\beta} = (\beta_1, ..., \beta_T)$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$ are temporal effects which account for the temporal structure of the data, and η_{it} is an space-time interaction which accounts for the possibility of different temporal trends in different areas. This model can be thought as the convolution model outlined in Section 2.3.4 in a spatio-temporal context since the structured and unstructured temporal effect terms are added to this model.

Here μ can be assigned by a non-informative prior and the two sets of unstructured independent main effects are assumed to have multivariate normal prior i.e. $\boldsymbol{\alpha} \sim \text{N}(\mathbf{0}, \sigma_\alpha^2 \boldsymbol{I})$ and $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I})$. The spatially structured random effects $\boldsymbol{\phi}$ are modelled by the intrinsic CAR model introduced in Section 2.3.4. Furthermore, the structured temporal random effects $\boldsymbol{\delta}$ are modelled by a first order random walk given by $\delta_1 \sim \text{N}(0, \sigma_\delta^2)$ and $\delta_t | \delta_{t-1} \sim \text{N}(\delta_{t-1}, \sigma_\delta^2)$ for $t = 2, ..., T$. There are four possible combinations to assign the prior distribution for the space-time interaction term η_{it}. The first is the interaction between unstructured spatial effects $\boldsymbol{\alpha}$ and unstructured temporal effects $\boldsymbol{\beta}$, in which case the interaction term η_{it} follows an independent prior distribution for all i and t. This interaction term accounts for unexplained effects which do not have any spatial or temporal structure. The second is an interaction between unstructured spatial effects $\boldsymbol{\alpha}$ and structured temporal effects $\boldsymbol{\delta}$, and then the interaction terms $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{iT})$ can be modelled by separate random walks for each unit. For an interaction between structured spatial effects $\boldsymbol{\phi}$ and unstructured temporal effects $\boldsymbol{\beta}$ the interaction terms $\boldsymbol{\eta}_t = (\eta_{1t}, ..., \eta_{nt})$ follow a CAR model for each time period. Finally, an interaction between structured spatial effects $\boldsymbol{\phi}$ and structured temporal effects $\boldsymbol{\delta}$ follows the conditional distribution of η_{it}, i.e. $\eta_{it} | \eta_{-it} \sim \text{N}(\psi_{it}, \sigma^2_{\eta(it)})$. The

mean and the variance can be computed as given

$$\begin{aligned} \psi_{it} &= \frac{1}{2}(\eta_{i,t-1} + \eta_{i,t+1}) + \frac{\sum_{j=1}^{n} w_{ij}\eta_{jt}}{\sum_{j=1}^{n} w_{ij}} + \frac{\sum_{j=1}^{n} w_{ij}(\eta_{i,t-1} + \eta_{i,t+1})}{2\sum_{j=1}^{n} w_{ij}} \\ \sigma^2_{\eta(it)} &= \frac{1}{2\sum_{j=1}^{n} w_{ij}}, \end{aligned} \tag{2.4.3}$$

where $\boldsymbol{W}$ is an $n \times n$ neighbourhood matrix with $w_{ij} = 1$ if areas (i, j) are neighbours, otherwise $w_{ij} = 0$. More details about the neighbourhood matrix are given in Section 2.3.2. To select an appropriate structure for the interaction term, it depends on the structure of the remaining variation in the data after the spatial and temporal main effects have been estimated. One drawback of this model is the linear predictor contains two variables per area, two variables per time period, and one interaction term per data point. Therefore five different sources of variation need to be estimated which can present challenges both in terms of computation and interpretability.

2.4.3 Ugarte Model

Ugarte et al. (2012) dropped some parameters from the model proposed by Knorr-Held (2000), as the structured and unstructured main effects are combined to a single parameter for both spatial and temporal effects. The model takes the form

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(e_{it}\theta_{it}) \qquad i = 1, ...n, \qquad t = 1, ..., T \\ \log(\theta_{it}) &= \mu + \phi_i + \delta_t + \eta_{it}, \end{aligned} \tag{2.4.4}$$

where μ is the global intercept, ϕ_i represents the spatial effect at area i, δ_t represents the temporal effect during time t, and η_{it} denote a space-time interaction. To assign prior distributions to each parameter, μ can be modelled by a non-informative prior. The spatial random effects $\boldsymbol{\phi}$ follow a CAR prior proposed by Leroux et al. (2000) which contains a spatial autocorrelation parameter to control the level of spatial smoothness in the data. The temporal random effects $\boldsymbol{\delta}$ are modelled by the first order random walk as described above. Finally the interaction effects $\boldsymbol{\eta}$ can be modelled by a normal prior with mean zero and a precision matrix which can be computed by the Kronecker

product of the precision matrices for the spatial and temporal effects. This model is simpler than the one proposed by Knorr-Held (2000) since the formula contains fewer variance parameters to be estimated.

2.4.4 Rushworth model

Rushworth et al. (2014) proposed a spatio-temporal model which has fewer parameters than the Ugarte model and is given by

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(e_{it}\theta_{it}) \qquad i = 1, ...n, \qquad t = 1, ..., T \\ \log(\theta_{it}) &= \mu + \eta_{it}, \end{aligned} \tag{2.4.5}$$

where η_{it} denote space-time random effects which are represented by a Gaussian Markov random field (GMRF) prior distribution. They assume that observations which are close together in time or space are likely to be correlated, and thus adopt multivariate first order autoregressive structure. The joint prior distribution for $\boldsymbol{\eta}_1 = (\eta_{11}, ..., \eta_{n1})$ is given by $\boldsymbol{\eta}_1 \sim \text{N}(\mathbf{0}, \tau^2 \boldsymbol{Q}(\rho, \boldsymbol{W})^{-1})$ where the precision matrix $\boldsymbol{Q}(\rho, \boldsymbol{W})$ is modelled via the CAR model proposed by Leroux et al. (2000), which is given by $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1 - \rho)\boldsymbol{I}$. Here $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is the $n \times 1$ vector of ones. The conditional distribution for the random effects at time t is as follows:

$$\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1} \sim \text{N}(\alpha \boldsymbol{\eta}_{t-1}, \tau^2 \boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}) \quad t = 2, ..., T. \tag{2.4.6}$$

Here, the temporal autocorrelation is induced via the mean $\alpha \boldsymbol{\eta}_{t-1}$, while the spatial autocorrelation is induced through the precision matrix $\boldsymbol{Q}(\rho, \boldsymbol{W})$.

2.5 Spatially rescaled models

The simplest situation of spatial misalignment people have considered in disease mapping is multiple scale modelling, where data are available at two (or more) nested spatial scales. Ugarte et al. (2016) modelled data on the fine spatial scale of municipalities, but aimed to make inference about larger units called provinces. The motivation is that each province has different health care implementation policies, and they are

aiming to see if these differences have an impact on disease risk. There is a many to one mapping between municipalities and provinces, so that each municipality is located within exactly one province. They modelled repeated municipality level count data on brain cancer incidence over time in Spain, and included spatial and temporal main effects as well as a space-time interaction. They linked the two spatial scales by including a random effect for province, which like the municipality level spatial effects, were modelled with a conditional autoregressive prior to allow for spatial autocorrelation. They showed by simulation that their model performs better than single level models in this two-level data setting, and then identified high risk provinces in Estella and Pamplona in the real data application. Although this model does have two spatial scales, the unique data really only exist at the small level, as the province level data are simple aggregates of the municipality level data.

Aregay et al. (2017) extended this idea by developing multiple spatial scale modelling that accounts for spatial autocorrelation at the lower (municipality) and higher (province) levels simultaneously using shared and separate random effect components. In the simplest case, a joint convolution model at the lower and higher levels is considered, and the linkage between these two levels is incorporated in the model by including a shared random effect of the higher level. The second proposed model includes an extra spatially structured random effect at the lower level. They compared these models to the case where they simply fitted separate convolution models at both levels. Finally, they also considered a model that obtains disease risk at the higher level by aggregating over the lower level estimated risks. The comparison of these models was made via a simulation study, and the results indicate that the shared random effect model with the extra random effect at the finer level performs best in regard to the bias and MSE of the estimated risks. This is because the lower level inherits a common characteristic from the higher level via the shared random effect, leading to better estimation. These proposed models were applied to Georgia oral cancer data, and the results were similar to the simulation study. The shared random effect model has the best performance at the higher level in terms of model fit, while at the lower level there is not much difference in the estimates.

Both these papers tackled the problem of spatial misalignment, but in the simplest case of there being exact nesting between the two spatial scales. If there is not exact nesting then we need to consider an alternative approach. The Log Gaussian Cox process (LGCP) is a commonly used model for spatial point patterns analysis. Li et al. (2012a) proposed modelling aggregated disease case count data with a spatially continuous LGCP. The continuous risk surface is approximated by a piecewise constant surface evaluated on regular grid squares. They assumed a hierarchical structure, where the first level involved a Poisson process with random intensity, and the second level had an intensity function drawn from a Gaussian Markov random field. They made statistical inference on the counts in the intersection areas between regions and grid squares. To perform inference, a data augmentated MCMC algorithm was used to produce samples for both model parameters, and the disease counts in the intersection areas. The disease counts were sampled from a multinomial distribution with the weight proportional to the size of the intersection areas. They compared the proposed model to the commonly used BYM model via a simulation study, and the results indicated that the proposed LGCP model outperformed the BYM since it produced more accurate disease risk maps in terms of MSE and was also better at identifying areas of abnormally high risk as measured by ROC curves. However, one drawback of this approach is that the expected disease counts and covariates are ignored in the multinomial steps. Additionally, I still obtain disease risk estimates in areas where no people live which is not realistic.

Li et al. (2012b) developed their methodology from the previous paper by addressing the problem of spatial modelling when the area boundaries change. They aimed to make inference on systemic lupus erythematosus (SLE) data for 40 years to 2007 and identified high risk areas in Toronto, Canada. The locations in space and time for individual cases are assumed to come from a spatio-temporal inhomogeneous Poisson process, while the random spatial risk surface is modelled via an LGCP. The disease cases in each grid cell were estimated by a generalised linear mixed model, with the spatial random effects approximated by a Gaussian Markov random field. In addition, the offset parameter was based on the population size in each grid cell, age group effect and the variation in risk over time. The proposed model was compared to the more

established BYM model via a simulation study, and was shown to perform better in terms of MSE. However, they still estimated disease risk for the grid squares with zero population.

Diggle et al. (2013) aimed to provide a map that estimates the spatial variation in the risk of lung cancer in the Castile-La Mancha in Spain. They partitioned each region of interest into subregions to make inference, however the subregions have different sizes and shapes. Therefore, the inference is made on regular grid cells via an LGCP model and a data augmentation approach as the previous approaches did.

Taylor et al. (2018) proposed spatial and spatio-temporal models that aimed to make continuous inference at the fine grid level based on aggregated disease count data where disease counts relate to (i) non-overlapping areas, (ii) overlapping areas, or (iii) areas with unknown boundaries such as usage of healthcare facilities. Multinomial sampling is again used to estimate disease cases at the grid level. The probability of each disease case in each region occurring in the intersection area is based on the size of the intersection area and covariate information, and different specifications are used depending on the boundary type. Inference was carried out via a data augmentation approach and an MCMC algorithm. They applied their method to three datasets, primary biliary cirrhosis cases between 1987 to 1994 in Newcastle upon Tyne, malaria over a two year period in Namibia with overlapping catchment areas, and general election results from 2010 and 2015 in Manchester and surrounding areas. The aim in the latter was to predict the results in 2020 when the electoral boundaries have changed. However, there are a few limitations in this study. Firstly, the expected disease counts are not included in the probabilities in the multinomial steps, whereas the estimated disease counts in each grid cell should depend on the population size that live there. Secondly, they estimated the disease risks in areas where no people live, for example mountains and fields. Moreover, they did not assess the risk estimation accuracy via a simulation study. Finally, they used an LGCP in their study which is more computationally demanding than the CAR model used here.

Therefore, in this thesis I aim to solve these issues by estimating disease risks on regular grid squares and hence create an approximate spatially continuous risk surface over the Greater Glasgow and Clyde Health Board. The disease counts at the grid square level are estimated via multinomial step with probabilities based on the expected counts and the sizes of areas of intersection between grid squares and regions. I propose two methods to achieve this goal; multiple imputation and data augmentation approaches. Then finally I extend these methods to estimate the spatio-temporal variation in disease risk.

Chapter 3

Spatial modelling for respiratory disease risk at the areal unit level

3.1 Introduction

Disease risk varies over space and time, and poverty and deprivation are significant factors that drive the spatial variation that can be observed in disease risk, with more affluent areas normally having lower levels of disease risk, while more deprived areas usually exhibit higher risk levels (McCartney, 2012). Disease mapping methods are most commonly used to estimate disease risk over space and time, and therefore areas of high or low risk can be identified. One aim in doing this is to estimate health inequalities, which can be defined as the variation in health risk between different social groups and population areas (Murray et al., 1999), and refers to the unfair and avoidable differences in people's health. These inequalities are mainly based on socio-economic factors, for example education, income and wealth (Jack et al., 2019). There have been many previous studies, such as Levin and Leyland (2006), Ellis and Fry (2010) and Marmot et al. (2010), focusing on health inequalities in large scale areas e.g. between countries. In this chapter I will explore the inequality in respiratory hospital admissions in the Greater Glasgow and Clyde Health Board at a small area level known as intermediate zones using an existing spatial correlation model called the Leroux conditional autoregressive (CAR) model (Leroux et al., 2000). These results will motivate the development of the novel methodology in the following chapters.

The remainder of this chapter is organised as follows. Section 3.2 introduces the dataset which will be used in this chapter and throughout this thesis. Then Section 3.3 outlines a spatial model used to estimate disease risk at the intermediate zone level, and also presents the results from the model. Finally Section 3.4 discusses the results and some drawbacks of this study.

3.2 Data

The study region is the Greater Glasgow and Clyde Health Board, which contains the largest city in Scotland (Glasgow) and the surrounding areas including areas of East Dunbartonshire, East Renfrewshire, Glasgow City, Inverclyde, Renfrewshire and West Dunbartonshire. These areas are called council areas and Table 3.1 presents the population in each council area. The Health Board area is split into 257 administrative units called intermediate zones (IZ) which are presented in Figure 3.1. These IZs have a median area of 119 hectares, with a maximum of 11,300 and a minimum of 20 hectares, while the median population is 4,306 with a maximum of 9,008 and a minimum of 1,321 (Scottish Government, 2019). The disease data are two-year total counts of the numbers of hospital admissions with a primary diagnosis of respiratory disease for the years 2015 to 2016 in each IZ and are collected from all hospitals in the health board (35 hospitals). The respiratory disease data are defined using the International Classification of Diseases Volume 10 (ICD10) codes (J00:J99, R09.1). The total number of disease counts for all IZs is 51,271 cases with a median of 188 and range is between 50 to 530 cases. Note that these disease data will be used throughout this thesis but in Chapter 6 will used the data from 2013 to 2016.

The disease data, $\boldsymbol{Y}(\mathcal{A}) = [Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$, are obtained from the Scottish Statistics website https://statistics.gov.scot. These data are denoted by $\boldsymbol{Y}(\mathcal{A}_i) = [Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$, and are the numbers of hospital admissions for respiratory disease for each area $\mathcal{A}_i$. The expected values, $\boldsymbol{e}(\mathcal{A}_i) = [e(\mathcal{A}_1), ..., e(\mathcal{A}_n)]$, are the expected hospital admission numbers for each area $\mathcal{A}_i$ computed to adjust for varying population sizes and demographic structures in each IZ. The expected values can be computed via

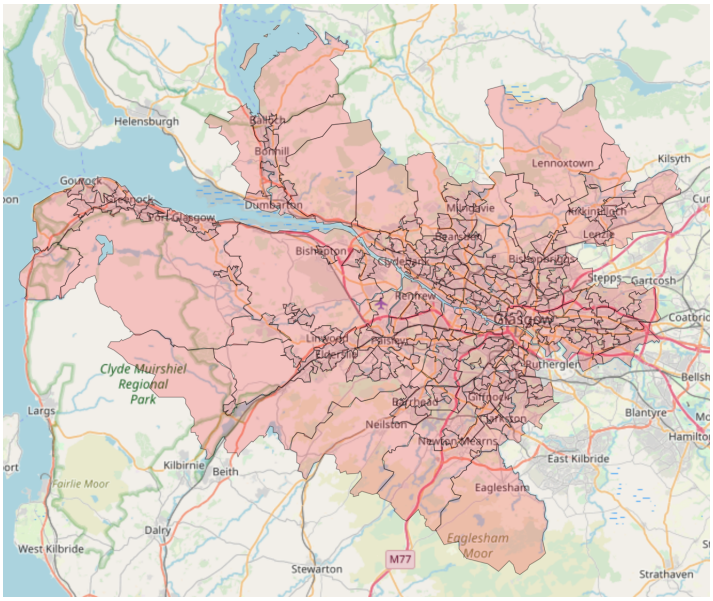


Figure 3.1: The intermediate zones of the Greater Glasgow and Clyde Health Board.

Table 3.1: Population estimates in the Greater Glasgow and Clyde Health Board for 2016.

Area	Population
East Dunbartonshire	107,540
East Renfrewshire	93,3810
Glasgow City	615,070
Inverclyde	79,160
West Dunbartonshire	89,860

Source: Population Estimates (Current Geographic Boundary), available at https://statistics.gov.scot/data_home.

indirect standardisation based on age and sex specific disease rates for the whole of Scotland. In other words, the expected disease count for area $\mathcal{A}_i$ can be calculated by first constructing a set of B strata of the population in each area based on age and sex. Then compute the expected disease counts via $e(\mathcal{A}_i) = \sum_{b=1}^{B} N_b(\mathcal{A}_i) r_b$, where $N_b(\mathcal{A}_i)$ is the population in area $\mathcal{A}_i$ in strata b and r_b is the average disease rate for strata b in Scotland. The simplest measure of disease risk is the standardised incidence ratio (SIR), which can be calculated by $\text{SIR}(\mathcal{A}_i) = Y(\mathcal{A}_i)/e(\mathcal{A}_i)$. An SIR value greater than 1 indicates that there is a higher disease incidence rate within the areal unit than the average over Scotland. In contrast a value less than 1 indicates a lower incidence rate than the average over Scotland.

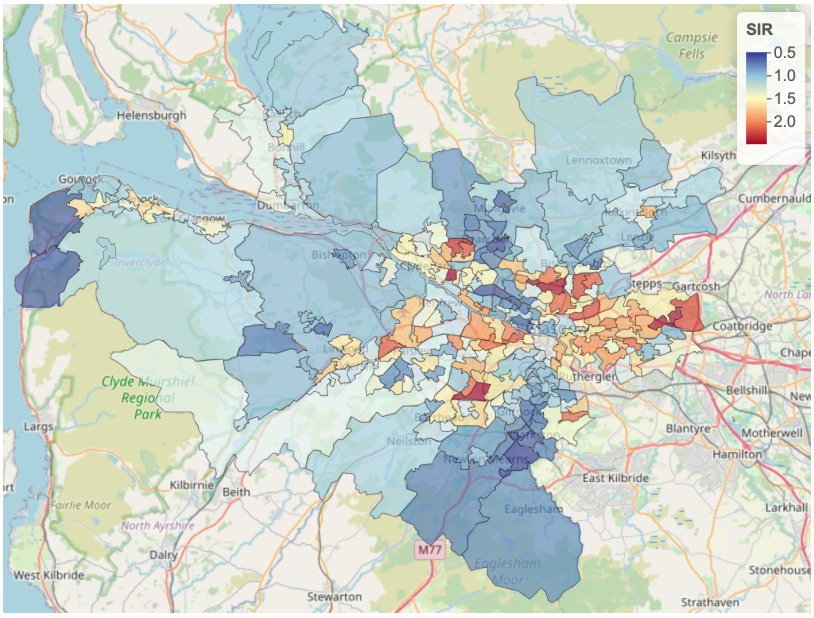


Figure 3.2: The standardised incidence ratio (SIR) for respiratory disease risk across the Greater Glasgow and Clyde Health Board for the years 2015 - 2016.

Based on the two years disease data, the total disease counts for the Greater Glasgow and Clyde Health Board is 51,271 with the maximum of 530 and minimum of 50, while median disease count is 188. In addition, the median SIR for the whole study region is 1.27 with the maximum of 2.48 and minimum of 0.50. Figure 3.2 presents the spatial map of the SIR for each IZ across the Greater Glasgow and Clyde Health Board in the years 2015 to 2016. The areas with higher SIRs are Clydebank, Paisley, and Barrhead, which are the poorer areas in Glasgow. In contrast the areas with lower SIRs are Giffnock, Clackston, Eaglesham, Bishopton and Bearsden, which are the wealthy areas. These results suggest that the wealthier areas tend to have lower SIRs than the less wealthy areas.

There are however some disadvantages of the SIR as the measure of risk. When the studied disease is rare or the populations in some areas are very small, some areas may have low values of the expected disease count $e(\mathcal{A}_i)$ which may result in very unstable SIR values in some areas as the SIR is a ratio. Furthermore the SIR is computed independently for each area, and therefore it does not take account of the spatial autocorrelation that might be present in the data. It is therefore more common to take a hierarchical Bayesian modelling approach to estimate disease risk. This approach often extends a Poisson GLM to allow for spatial autocorrelation, which is included

via a set of random effects. These random effects borrow strength from geographically nearby areas to improve the accuracy of estimation.

3.3 Estimating disease risk at the areal unit level

In this section I will estimate respiratory disease risk at the areal unit level across the Greater Glasgow and Clyde Health Board using the commonly used Leroux CAR model (Leroux et al., 2000).

3.3.1 Spatial modelling

The spatial model used for the disease count response data is typically a Poisson GLM with random effect and is outlined as follows:

$$\begin{aligned} Y(\mathcal{A}_i) &\sim \text{Poisson}[e(\mathcal{A}_i)R(\mathcal{A}_i)] \qquad i = 1, ..., n, \\ \ln[R(\mathcal{A}_i)] &= \mathbf{x}(\mathcal{A}_i)^\top \boldsymbol{\beta} + \phi(\mathcal{A}_i), \end{aligned} \tag{3.3.1}$$

where $Y(\mathcal{A}_i)$ and $e(\mathcal{A}_i)$ respectively denote the number of observed and expected respiratory disease cases in area $\mathcal{A}_i$. $R(\mathcal{A}_i)$ is the disease risk in area $\mathcal{A}_i$ which can be estimated via the covariate information, $\mathbf{x}(\mathcal{A}_i)^\top \boldsymbol{\beta}$, and a set of random effects, $\boldsymbol{\phi}(\mathcal{A}) = [\phi(\mathcal{A}_1), ..., \phi(\mathcal{A}_n)]$. Here the covariates are not included in this study so $\mathbf{x}(\mathcal{A}_i)^\top \boldsymbol{\beta} = \beta_0$ since the grid level analysis presented in subsequent chapters shows they are not well estimated. The intercept parameter β_0 is assigned a normal prior distribution with mean zero and variance 100,000, i.e. $\beta_0 \sim \text{N}(0, 100,000)$. The random effects are used to account for the spatial autocorrelation that might be present in the data. These random effects are typically modelled via a conditional autoregressive (CAR) model. Here I use the CAR prior proposed by Leroux et al. (2000) which is given by

$$\begin{aligned} \phi(\mathcal{A}_i)|\boldsymbol{\phi}(\mathcal{A}_{-i}) \sim \; & N\left(\frac{\rho \sum_{k=1}^{n} w_{ik}\phi(\mathcal{A}_k)}{\rho \sum_{k=1}^{n} w_{ik} + (1-\rho)}, \frac{\tau^2}{\rho \sum_{k=1}^{n} w_{ik} + (1-\rho)}\right), \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b), \\ \rho &\sim \text{Uniform}(0, 1), \end{aligned} \tag{3.3.2}$$

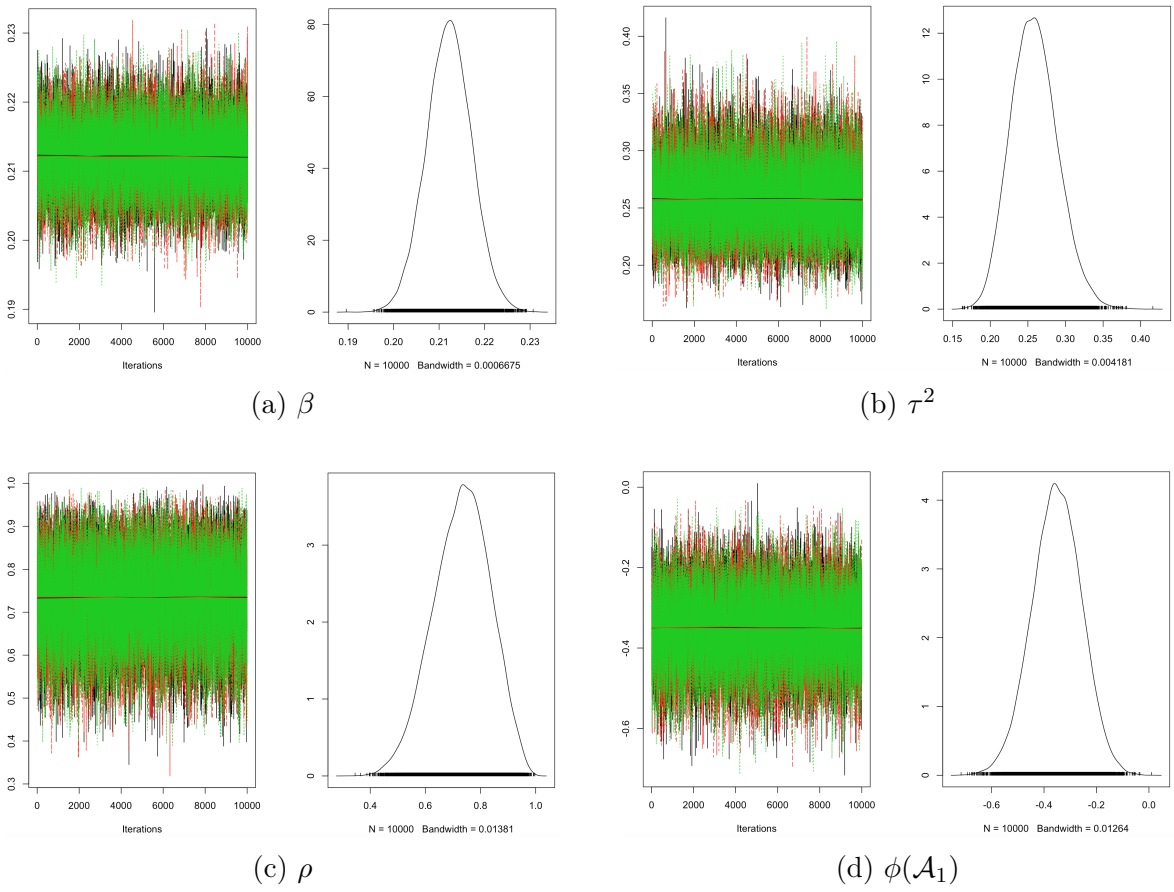
where $\boldsymbol{\phi}(\mathcal{A}_{-i}) = [\phi(\mathcal{A}_1), ..., \phi(\mathcal{A}_{i-1}), \phi(\mathcal{A}_{i+1}), ..., \phi(\mathcal{A}_n)]$. Here ρ denotes the spatial autocorrelation between these random effects and is controlled by an $n \times n$ neighbourhood matrix $\boldsymbol{W}$, which is described in Section 2.3.3, with $w_{ij} = 1$ if areas A_i and A_j share a common border and $w_{ij} = 0$ otherwise. A value of $\rho = 1$ indicates strong spatial autocorrelation (the intrinsic CAR model (Besag et al., 1991)), while $\rho = 0$ indicates that the random effects are completely independent ($\phi(\mathcal{A}_i) \sim \mathrm{N}(0, \tau^2)$). The variance parameter τ^2 controls the level of variation between the random effects, which is assigned an inverse-Gamma prior distribution with a shape parameter a and a scale parameter b.

3.3.2 Results

Model inference is performed by using an MCMC algorithm via a combination of Gibbs sampling and Metropolis-Hasting steps. The MCMC algorithm is implemented using the CARBayes package (Lee, 2013) in R (R Core Team, 2014). The model is run three times in order to generate three independent Markov chains, and each chain is run for 200,000 iterations with 50,000 as burn-in iterations and thinned by 15. This gives 30,000 remaining samples for overall inference, with 10,000 samples for each chain.

Convergence diagnostic

To assess convergence of the Markov chains, the simplest way is to draw a traceplot of the posterior samples for each parameter, and the convergence is presented when the samples show no clear pattern (e.g. always increasing) in such a plot. In general, the convergence diagnostic should be applied for every model parameter however, this is infeasible in practice since there are a large number of random effects. Therefore only selected parameters which are (β, τ^2, ρ) and ten random effects ($\phi(\mathcal{A}_i)$) are undertaken. Figure 3.3 presents trace plots of the posterior samples for (β_0, τ^2, ρ) and random effect for area $\mathcal{A}_1$, and each chain is represented in different colours. The figure shows the MCMC chains have converged since there is no change in a mean or variance of the posterior samples among the three chains. Note that, the convergence of random effects for ten areas has been checked, however it is presented only one area since the



(a) β (b) τ^2

(c) ρ (d) $\phi(\mathcal{A}_1)$

Figure 3.3: Trace plots of the MCMC samples from each parameter.

plots look very similar.

An additional diagnostic is the method of Gelman-Rubin (Gelman and Rubin, 1992), used to assess the convergence for multiple chains. Gelman et al. (2013) suggest that a value less than 1.1 indicates good mixing of the chain. The Gelman-Rubin statistics for selected parameters are all less than 1.1 with a maximum value of 1.02. Typically the results show that the posterior samples are well mixed.

Sensitivity analysis

Sensitivity analysis is carried out in order to ensure that the posterior distribution is not driven by a choice of hyperpriors. Here, three scenarios with different hyperpriors for the variance of random effects τ^2 are operated as follows:

1. Scenario 1 - $\tau^2 \sim$ Inverse-Gamma$(1, 0.01)$
2. Scenario 2 - $\tau^2 \sim$ Inverse-Gamma$(0.01, 0.01)$

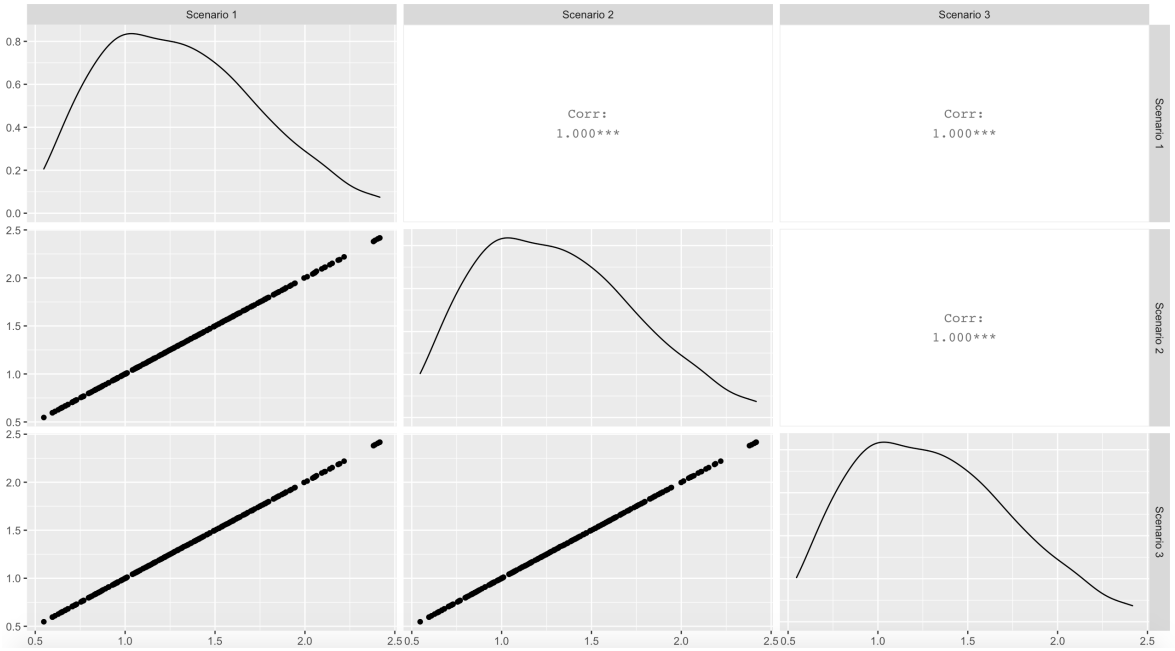


Figure 3.4: Correlation plots of estimated risks between three MCMC chains.

3. Scenario 3 - $\tau^2 \sim$ Inverse-Gamma$(0.05, 0.0005)$

These choices of hyperparameters are selected regarding to Lee (2020), Rodrigues and Assunção (2012), and Law (2016). Figure 3.4 presents the correlation plots between the risk estimates across scenarios. The plots show that the risk estimates lie on the straight line, which indicates all three scenarios producing very similar estimates of the risk. These results indicate that different hyperpriors do not affect the posterior distribution, therefore the posterior samples of one scenario are used for the model inference.

Posterior predictive check

Posterior predictive check is a goodness of fit assessment criteria to check if the model is fitted the data well which was introduced by Rubin (1984) and was extended by Gelman et al. (1996). The idea is to compare simulating replicated data under the fitted model to the observed data. Intuitively, if the model specifications are appropriate, the simulated data and the observed data should be the same. Figure 3.5 presents the correlation plot between the observed data and the average of simulated data for each areal unit. The plot indicates that both data sets are similar since the data lie on the straight line with correlation coefficient of 0.999.

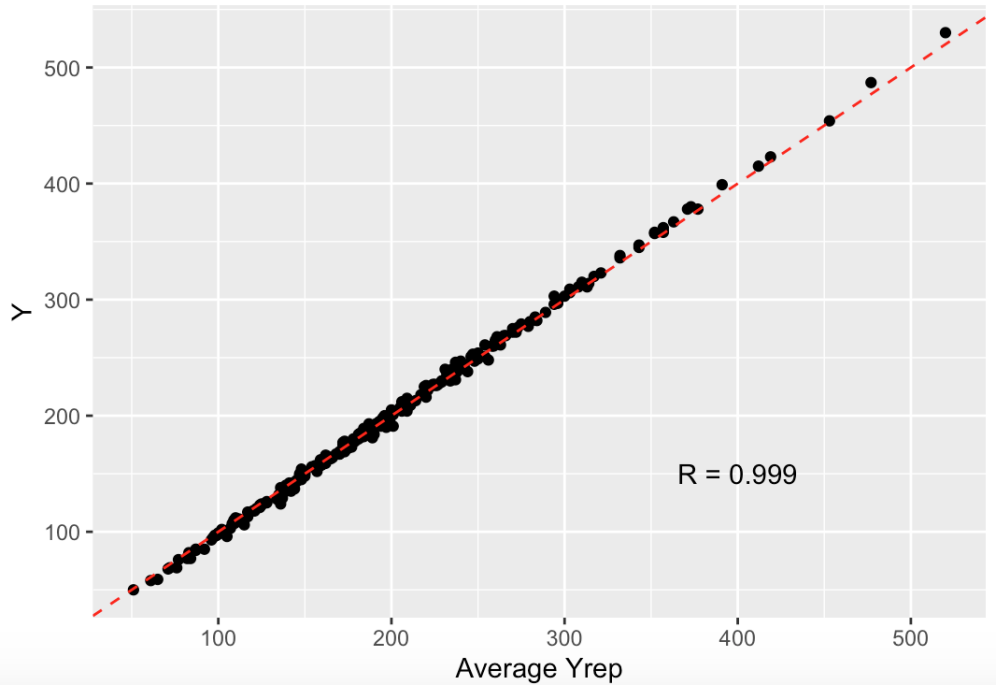


Figure 3.5: Posterior predictive model checking.

Main results

Point estimates of the parameters are taken from the median of the posterior distribution of each parameter, and the 95% credible interval is taken from the 2.5 and 97.5 percentiles of the posterior distribution. The estimate of the spatial autocorrelation $\hat{\rho}$ is 0.74 with the 95% credible interval (0.52, 0.92). The credible interval is not near zero, therefore there is sizeable spatial association present in the data. In addition the estimate of the random effect's variance $\hat{\tau}^2$ is 0.26 with 95% credible interval (0.20, 0.32), suggesting sizable variation in thedisease risk (τ^2 is on the log scale). Figure 3.6 shows the estimated respiratory disease risk from the model introduced in Section 3.3.1. The spatial estimated risk map shows a similar spatial pattern to the SIR map presented in Figure 3.2. To quantify this similarity, Figure 3.7 illustrates the correlation of the SIRs in each IZ and the estimated risks from Model (3.3.1), which is 0.999, therefore this spatial model and SIR approach produce very similar risk estimates. However, the variation of the risk estimates is slightly smaller than the SIRs (0.641 vs 0.665) as seen in Figure 3.8. The numbers in red on the top of each boxplot are the interquartile ranges (IQR), which is used to measure the variability in the disease

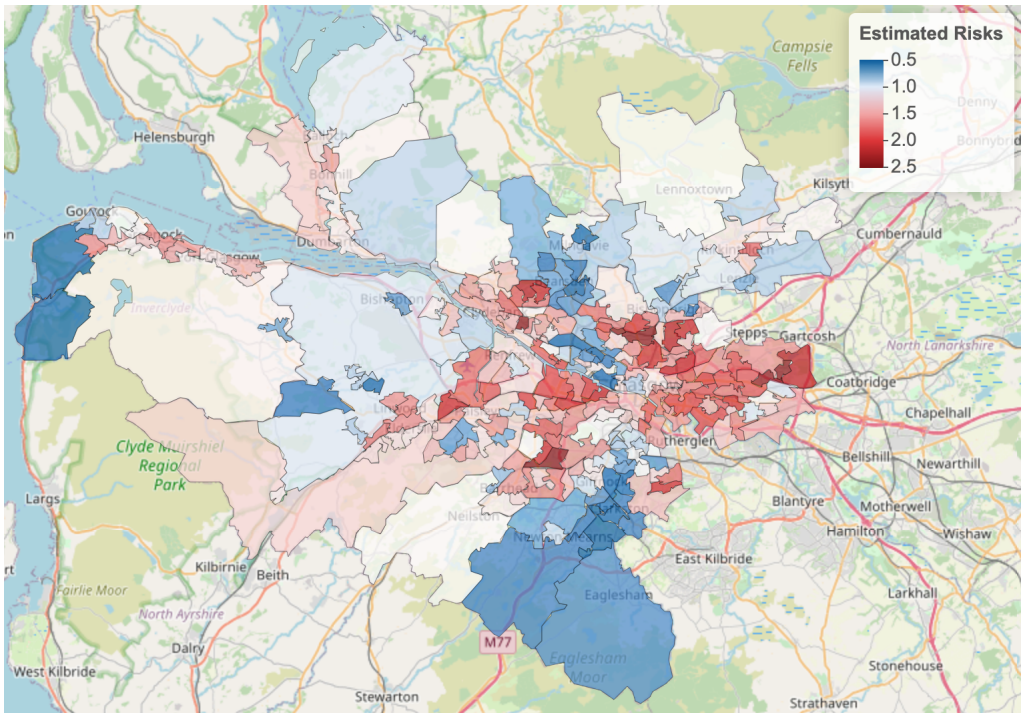


Figure 3.6: The estimated respiratory disease risk across the Greater Glasgow and Clyde Health Board.

risk and it can be computed as the difference between the upper and lower quartiles. The results collectively suggest that the higher risk areas are located both in the east and west of Glasgow city centre. These areas are among the most deprived areas in Glasgow e.g. Clydebank and Paisley. In contrast, the areas with lower risks which are wealthier areas such as Bearsden, Clarkston and Eaglesham are mainly located in the north and south of the city centre.

3.4 Conclusion

In this chapter the hierarchical Bayesian model proposed by Leroux et al. (2000) has been used to estimate the respiratory disease risk across the Greater Glasgow and Clyde Health Board. This model has a parameter to control the level of the spatial autocorrelation between the random effects. Therefore, a strong spatial smoothness is not necessary to be assumed. Hence, this model is suitable for the spatial data with all levels of correlation, and in this study it is moderate to strong correlation ($\hat{\rho} = 0.74$). Overall the areas with the higher disease risks are located on both the east and west of Glasgow city centre such as Clydebank and Paisley. These areas are among the most

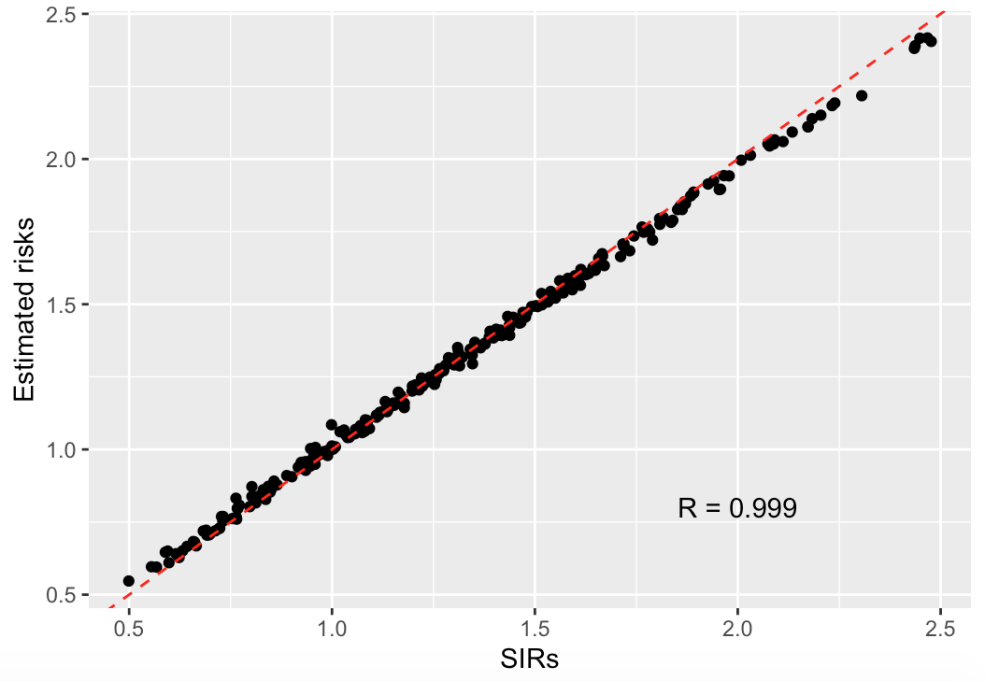


Figure 3.7: Scatter plot between SIRs and the estimated disease risks from the spatial model.

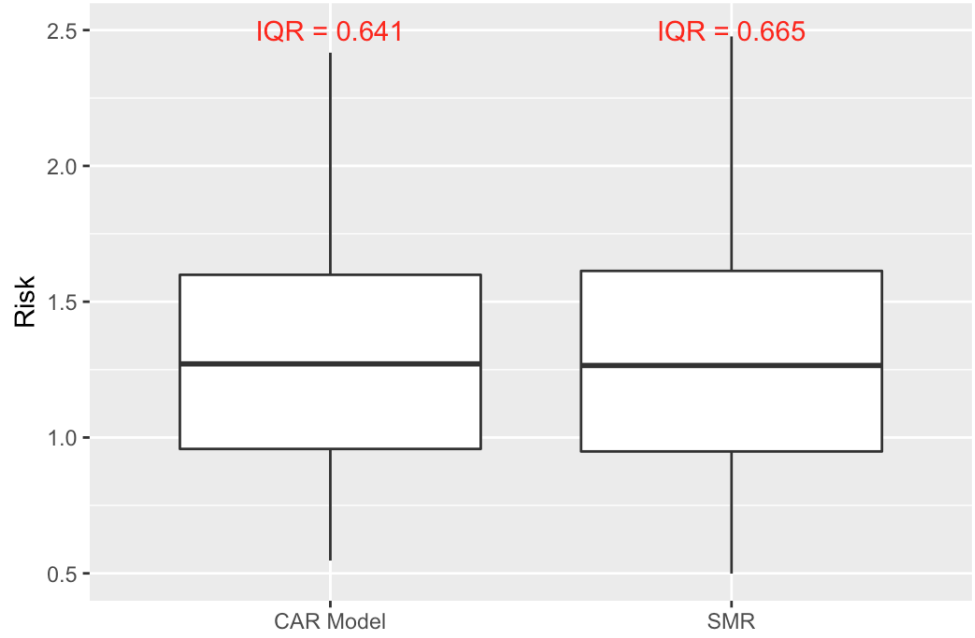


Figure 3.8: Boxplots of estimated respiratory disease risks using SIR approach and spatial modelling approach.

deprived areas in Glasgow. In contrast, the areas with the lower risks are in the north of the city centre and also in the south such as Bearsden, Clarkston and Eaglesham which are wealthier areas. These results suggest that people living in the more deprived areas are more likely to be hospitalised for respiratory disease than those living in richer areas. However, a major limitation of this study is the risk estimates which are assumed constant within each IZ, which is not necessary true. Therefore in the next chapters It would be overcome for these problems by proposing novel models which estimate the disease risk on a pseudo-continuous spatial surface using a grid based approach. This allows risk to vary within each IZ, thus providing a more accurate estimate of the spatially varying risk surface across the Greater Glasgow and Clyde Health Board.

Chapter 4

Grid level inference with multiple imputation

4.1 Introduction

The Modifiable Areal Unit Problem (MAUP) is a well known issue for people who analyse spatial data aggregated to non-overlapping areal units (Dark and Bram, 2007). There are two issues of concern related to MAUP: (i) the choice of unit is arbitrary and therefore when the units are changed, the data also change and hence change the results, (ii) when the set of areas are changed from one time period to the next period or more generally between two different data sets, we cannot directly compare the results because the units have been changed.

In the previous chapter, the existing method was presented in order to investigate health inequalities in respiratory disease hospital admission in the Greater Glasgow and Clyde Health Board. This method illustrated spatial disease risks at the areal unit level in disease maps using a hierarchical Bayesian model. There are however some drawbacks to this method in terms of estimating disease risk in each areal unit. For example, disease risk in each IZ is assumed to be constant, and it also estimates the risks in areas where nobody lives e.g. mountains, fields, which are not necessarily realistic.

Therefore this chapter will try to overcome these problems by developing methodology that uses areal unit level data to make grid square level inference, thus providing a common grid scale for inference regardless of the original areal units. Moreover, another benefit of this method is that as the grid squares become smaller the data and hence the inference will get closer to an individual level, moving away from the ecological fallacy (Wakefield and Salway, 2001). The ecological fallacy occurs when we make inference at the individual level based only on analyses of group or aggregated data in which those individuals belong.

The remainder of this chapter will be organised as follows. Section 4.2 introduces methodology of estimating grid level data based on areal unit level data, as well as the spatial model being used in this chapter and how to make an inference at the grid square level. Then a simulation study is carried out in Section 4.3 to test the performance for the proposed models. Section 4.4 applies respiratory hospital admissions data in the Greater Glasgow and Clyde Health Board to the selected models from Section 4.3. Finally, Section 4.5 summarises the main findings in this chapter and discusses the benefits and limitations of this methodology.

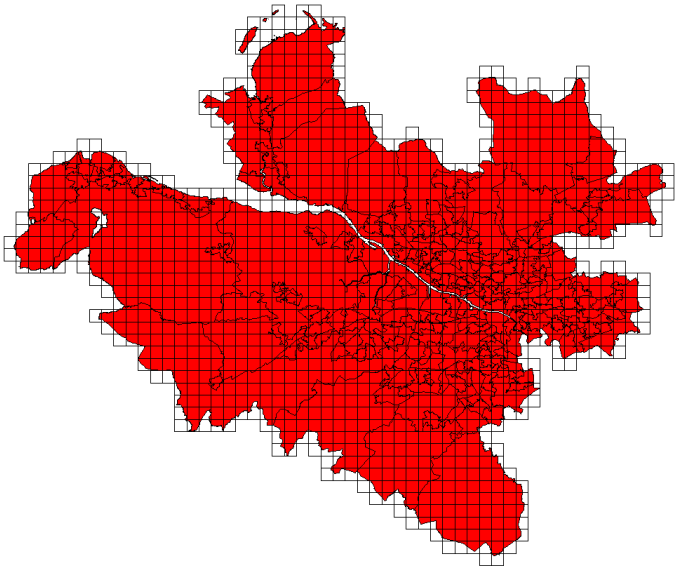


Figure 4.1: An example of grid squares over the Glasgow intermediate zone regions.

4.2 Methodology

4.2.1 The spatial grid

Suppose I have n regions $\boldsymbol{\mathcal{A}} = (\mathcal{A}_1, ..., \mathcal{A}_n)$, for which I have data $[Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$ and $[e(\mathcal{A}_1), ..., e(\mathcal{A}_n)]$, which denote the observed and expected numbers of disease cases respectively. In addition, $\mathbf{x}(\mathcal{A}_i) = [\mathrm{x}_1(\mathcal{A}_i), ..., \mathrm{x}_p(\mathcal{A}_i)]$ denotes a $p \times 1$ covariate vector for area $\mathcal{A}_i$ including $\mathrm{x}_1(\mathcal{A}_i) = 1$ for the intercept term. The expected number of disease cases for region $\mathcal{A}_i$, $e(\mathcal{A}_i)$ can be computed via indirect standardisation, to ensure that the observed disease cases, $Y(\mathcal{A}_i)$, are adjusted due to the populations in each region having different sizes and demographic structures e.g. age and sex. The aim is to provide grid square level inference based on areal unit data. Thus I overlay a grid of M squares $\boldsymbol{\mathcal{H}} = (\mathcal{H}_1, ..., \mathcal{H}_M)$ over the areas of interest, an example of which is shown in Figure 4.1. However, some grid squares have a zero population as they are areas of fields, mountains, lake, etc., so it is inappropriate to provide grid level inference in grid squares where no people live. Therefore grid squares with zero population have been removed in this study, and the remaining m grid squares with non-zero populations $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_1, ..., \mathcal{G}_m\} \subset \boldsymbol{\mathcal{H}}$ will be used for grid level inference, which is illustrated in Figure 4.2. Let $a(\cdot)$ denote area, so that $a(\mathcal{A}_i), a(\mathcal{G}_j)$, and $a(\mathcal{A}_i \cap \mathcal{G}_j)$ respectively denote the areas of region $\mathcal{A}_i$, grid square $\mathcal{G}_j$, and their intersection $\mathcal{A}_i \cap \mathcal{G}_j$. I wish to carry out grid level inference by fitting a spatial model to grid level data, $[Y(\mathcal{G}_j), e(\mathcal{G}_j), \mathbf{x}(\mathcal{G}_j)]$, which denote the unknown observed and expected disease counts for grid square $\mathcal{G}_j$ and their covariates. Then let $\tilde{P}(\mathcal{G}_j)$ denotes the population living in grid square $\mathcal{G}_j$, these data can be obtained from population density maps (Reis et al. (2017)), as shown in Figure 4.3. However, there is a problem with the grid squares that lie on the border of the map as illustrated in Figure 4.4, since some part of these grid squares is not included in region $\boldsymbol{\mathcal{A}}$. Let us consider two cases

1. If some part of the grid square that lies outside the map is uninhabitable (e.g. the sea), then the population in the whole grid square will come from region $\mathcal{A}$.

2. If some part of the grid square that lies outside the map is inhabitable (e.g. a neighbouring region of Glasgow) as show in Figure 4.4, then the population in that grid square will only partially come from region $\boldsymbol{\mathcal{A}}$. Hence, I have to make

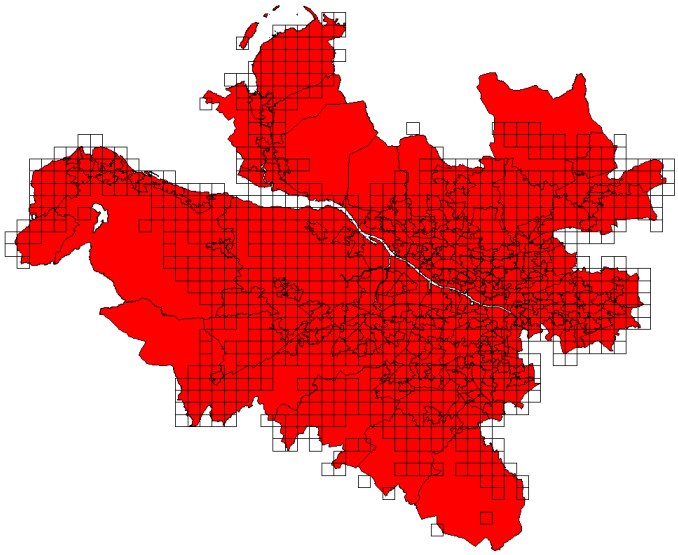


Figure 4.2: The Glasgow intermediate overlayed by grid squares with non-zero population.

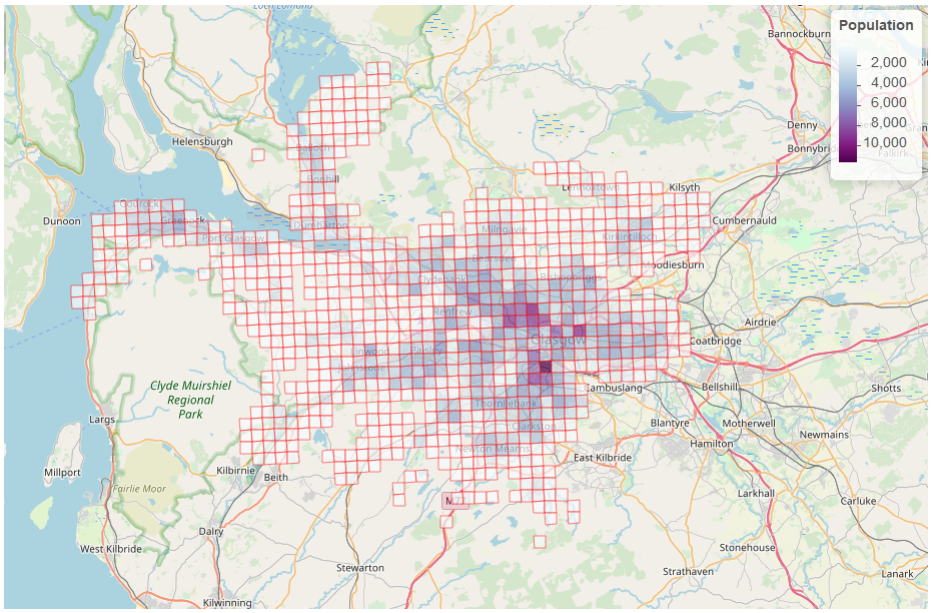


Figure 4.3: The population density at the grid square level overlaid on an OpenStreetMap.

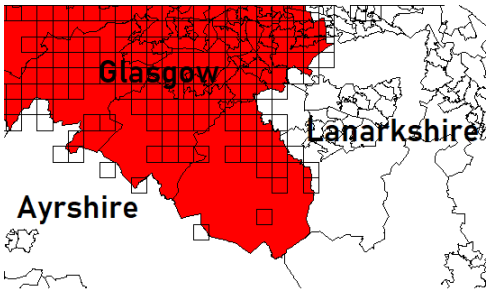


Figure 4.4: An example of grid squares which partly lie outside the Glasgow map.

an adjustment for the population in these grid squares to only relate to region $\boldsymbol{\mathcal{A}}$.

Therefore, under the assumption of a constant population density across a grid square, the population in grid square $\mathcal{G}_j$ is given by

$$P(\mathcal{G}_j) = \begin{cases} \tilde{P}(\mathcal{G}_j) & \text{case 1} \\ \left\lfloor \tilde{P}(\mathcal{G}_j) \frac{\sum_{i=1}^{n} a(\mathcal{A}_i \cap \mathcal{G}_j)}{a(\mathcal{G}_j)} \right\rceil & \text{case 2,} \end{cases} \tag{4.2.1}$$

so that in case 2 the population is reduced proportionally to the area of intersection with region $\mathcal{A}$. Here $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

4.2.2 Estimating grid level data $[e(\mathcal{G}_j), \mathrm{x}(\mathcal{G}_j)]$

Recall that I have areal unit level data $[e(\mathcal{A}_i), \mathbf{x}(\mathcal{A}_i)]$, and want to transform them to grid level data $[e(\mathcal{G}_j), \mathbf{x}(\mathcal{G}_j)]$. To compute the expected number of disease cases $e(\mathcal{G}_j)$, the total expected disease cases need to be reallocated from the n regions to the m grid squares. Note that I must have $\sum_{i=1}^{n} e(\mathcal{A}_i) = \sum_{j=1}^{m} e(\mathcal{G}_j)$ to keep the total expected counts as the same as the areal unit level. The expected disease cases for grid square $\mathcal{G}_j$ can be computed as $e(\mathcal{G}_j) = \sum_{i=1}^{n} e(\mathcal{A}_i \cap \mathcal{G}_j)$, the sum of the expected number of disease cases in the intersection areas between each region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$. Then assuming the expected counts are proportional to population density, the intuitive estimate of $e(\mathcal{G}_j)$ is

$$e(\mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{m} P(\mathcal{A}_i \cap \mathcal{G}_k)} e(\mathcal{A}_i), \tag{4.2.2}$$

which is a weighted average based on the population in the intersection area $\mathcal{A}_i \cap \mathcal{G}_j$.Here $P(\mathcal{A}_i \cap \mathcal{G}_j)$ is unknown, but it can be estimated based on the assumption of a common population density within each grid square. Therefore $P(\mathcal{A}_i \cap \mathcal{G}_j) = \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)} P(\mathcal{G}_j)$, the proportion of grid square $\mathcal{G}_j$ in region $\mathcal{A}_i$ multiplied by the population in grid square $\mathcal{G}_j$. Finally, substitute $P(\mathcal{A}_i \cap \mathcal{G}_j)$ into (4.2.2), so that

$$e(\mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{G}_j) \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} P(\mathcal{G}_k) \frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}} e(\mathcal{A}_i). \tag{4.2.3}$$

It its straightforward to show that the total number of expected disease counts in the grid level is equal to the areal unit level.

Covariates at the grid square level also need to be estimated. I consider three different types of covariate; continuous, count, and categorical data, and they must be estimated in different ways. The simplest approach for each type of covariate is given below

- Continuous data are a weighted average based on the population in the area of intersection, so that I have

$$\mathrm{x}(\mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} P(\mathcal{A}_q \cap \mathcal{G}_j)} \mathrm{x}(\mathcal{A}_i) = \sum_{i=1}^{n} \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)} \mathrm{x}(\mathcal{A}_i), \tag{4.2.4}$$

  which is the proportion of grid square $\mathcal{G}_j$ in region $\mathcal{A}_i$ multiplied by the covariate value in region $\mathcal{A}_i$ and assumes a common population density across the grid square.

- Count data are weighted in the same way as $e(\mathcal{G}_j)$, namely

$$\mathrm{x}(\mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{G}_j) \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} P(\mathcal{G}_k) \frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}} \mathrm{x}(\mathcal{A}_i). \tag{4.2.5}$$

- Categorical data modelled by binary indicators are estimated as in (4.2.4) except that the result is rounded to the nearest integer. Mathematically this can be

written as,

$$\mathrm{x}(\mathcal{G}_j) = \left\lfloor \sum_{i=1}^{n} \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{n} a(\mathcal{A}_k \cap \mathcal{G}_j)} \mathrm{x}(\mathcal{A}_i) \right\rceil, \tag{4.2.6}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

Note that few issues that occur for the count data are the sum of the data at the grid level is not necessarily equal to the areal unit level, i.e. $\sum_{j=1}^{m} \mathrm{x}(\mathcal{G}_j) \neq \sum_{i=1}^{n} \mathrm{x}(\mathcal{A}_i)$ and $\mathrm{x}(\mathcal{G}_j)$ is not necessarily an integer. To overcome these problems, I propose the following steps:

1. Round down all the values from (4.2.5), i.e. take the integer part of those values, $\kappa(\mathcal{G}_j) = \lfloor \mathrm{x}(\mathcal{G}_j) \rfloor$.

2. Denote the decimal part of the values from (4.2.5), $\mathrm{x}(\mathcal{G}_j)$, by $\pi(\mathcal{G}_j) = \mathrm{x}(\mathcal{G}_j) - \kappa(\mathcal{G}_j)$.

3. Sum the integer part from step 1, $\sum_{j=1}^{m} \kappa(\mathcal{G}_j)$, and compute the difference between the total number of the data at the areal unit level for the whole study region and the result from this step to find the number of data that have been lost in the transformation step, i.e. $d = \sum_{i=1}^{n} \mathrm{x}(\mathcal{A}_i) - \sum_{j=1}^{m} \kappa(\mathcal{G}_j)$.

4. Order the decimal parts, $\pi(\mathcal{G}_j)$, from step 2 and add one case to the $\kappa(\mathcal{G}_j)$ in grid squares with the d greatest decimal values, thus ensuring that $\sum_{j=1}^{m} \mathrm{x}(\mathcal{G}_j) = \sum_{i=1}^{n} \mathrm{x}(\mathcal{A}_i)$.

4.2.3 Methodology for estimating $Y(\mathcal{G}_j)$

Next, the disease cases at the grid level, $Y(\mathcal{G}_j)$, need to be estimated. Here, I propose estimating $Y(\mathcal{G}_j)$ by multiple imputation (Rubin, 2004), and then using these imputed data to fit a spatial model to estimate disease risk and covariate effects at the grid level. Multiple imputation generally consists three stages, the first stage is to generate multiple datasets of $Y(\mathcal{G}_j)$ based on $Y(\mathcal{A}_i)$ via multinomial sampling. Then fit a spatial model to each of the imputed datasets, and finally combine the results from the previous stage in order to make an inference.

Let us denote $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ as the number of disease cases in the intersection of $\mathcal{A}_i$ and $\mathcal{G}_j$. Then $Y(\mathcal{G}_j)$ is simply computed by

$$Y(\mathcal{G}_j) = \sum_{i=1}^{n} Y(\mathcal{A}_i \cap \mathcal{G}_j). \tag{4.2.7}$$

$Y(\mathcal{A}_i \cap \mathcal{G}_j)$ can be estimated by dividing the disease cases in region $\mathcal{A}_i$, $Y(\mathcal{A}_i)$ across the m grid square intersections $\{\mathcal{A}_i \cap \mathcal{G}_1, ..., \mathcal{A}_i \cap \mathcal{G}_m\}$ using a multinomial sampling step. Therefore given $Y(\mathcal{A}_i)$, I can generate $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ as follows

$$[Y(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y(\mathcal{A}_i \cap \mathcal{G}_m)] \sim \text{Multinomial}(n = Y(\mathcal{A}_i)|\omega_{i1}, ..., \omega_{im}). \tag{4.2.8}$$

Then combine $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ via (4.2.7) to estimate $Y(\mathcal{G}_j)$ for each grid square $\mathcal{G}_j$. However, I cannot complete the multinomial step if I do not know the probability (ω_{ij}) of each disease case in region $\mathcal{A}_i$ occurring in the intersection area $\mathcal{A}_i \cap \mathcal{G}_j$. Here, ω_{ij} will depend on two elements.

1. The size of the area of intersection between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$ relative to the other grid squares' areas of intersection,

$$\eta_{ij} = \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}.$$

2. The estimated number of disease events that we expect to observe in grid square $\mathcal{G}_j$,

$$\xi(\mathcal{G}_j) = e(\mathcal{G}_j)R(\mathcal{G}_j),$$

where $R(\mathcal{G}_j)$ is the grid level relative risk.

Then I combine η_{ij} and $\xi(\mathcal{G}_j)$ together, giving $\omega_{ij} \propto \eta_{ij}\xi(\mathcal{G}_j) = e(\mathcal{G}_j)R(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}$. I consider two possibilities for ω_{ij}:

Scenario 1: Constant risk across the region

Since I do not know the relative risk $R(\mathcal{G}_j)$ for any grid square, I assume equal risk across all grid squares by setting $R(\mathcal{G}_j) = 1$ for all j. This gives me:

$$\omega_{ij} = \frac{e(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} e(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}}, \tag{4.2.9}$$

and the denominator is used to ensure that $\sum_{j=1}^{m} \omega_{ij} = 1$ for all $i = 1, ..., n$.

Scenario 2: Risk estimates via kriging

I estimate $R(\mathcal{G}_j)$ at the grid square level for $j = 1, ..., m$ by applying spatial prediction methodologies to the standardised incidence ratio data at the areal unit level, $\hat{R}(\mathcal{A}_i) = Y(\mathcal{A}_i)/e(\mathcal{A}_i)$, because it is on the same risk scale as $R(\mathcal{G}_j)$. I consider this as essentially a geostatistical prediction problem where I have SIR data at the areal unit level $\{\hat{R}(\mathcal{A}_1), ..., \hat{R}(\mathcal{A}_n)\}$, and want to predict the SIR at the grid level $\{\hat{R}(\mathcal{G}_1), ..., \hat{R}(\mathcal{G}_m)\}$. Note that for this prediction exercise I assume that the SIR data in region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$ are located at their central points known as centroids $\mathfrak{a}_i$ and $\mathfrak{g}_j$ respectively. Here I use the method of kriging which is explained in Section 2.3.5 to predict the SIR values at locations $\mathfrak{g} = \{\mathfrak{g}_1, ..., \mathfrak{g}_m\}$ based on all SIR data at locations $\mathfrak{a} = \{\mathfrak{a}_1, ..., \mathfrak{a}_n\}$. I propose the geostatistical model as follows:

$$\hat{\boldsymbol{R}}(\mathfrak{a}) \sim \mathrm{N}(\mu\mathbf{1}, \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{a}}),$$

where $\hat{\boldsymbol{R}}(\mathfrak{a}) = \{\mathrm{SIR}(\mathfrak{a}_1), ..., \mathrm{SIR}(\mathfrak{a}_n)\}$ is the vector of SIR data at the n IZ centroids $\{\mathfrak{a}_1, ..., \mathfrak{a}_n\}$ and the centroids are purely land based. Here μ is the overall mean, and $\mathbf{1}$ is an $n \times 1$ vector of ones. Furthermore, $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{a}}$ is a covariance matrix which is specified by the exponential model, where $\boldsymbol{\theta}$ is the vector of nugget, sill and range parameters that need to be estimated. Then consider the joint geostatistical process at the n locations of IZs $\mathfrak{a} = \{\mathfrak{a}_1, ..., \mathfrak{a}_n\}$ and the m grid squares prediction locations $\mathfrak{g} = \{\mathfrak{g}_1, ..., \mathfrak{g}_m\}$:

$$\hat{\boldsymbol{R}}^* = \begin{bmatrix} \hat{\boldsymbol{R}}(\mathfrak{g}) \\ \hat{\boldsymbol{R}}(\mathfrak{a}) \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta_0 \boldsymbol{I} \\ \beta_0 \mathbf{1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{g}} & \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{ag}}^{\top} \\ \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{ag}} & \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{a}} \end{bmatrix} \right),$$

where $\hat{\boldsymbol{R}}(\mathfrak{g}) = [\hat{R}(\mathfrak{g}_1), ..., \hat{R}(\mathfrak{g}_m)]$. Additionally $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{g}} = \mathrm{Var}[\hat{\boldsymbol{R}}(\mathfrak{g})]$, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{a}} = \mathrm{Var}[\hat{\boldsymbol{R}}(\mathfrak{a})]$ are based on the exponential autocovariance. Furthermore $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathfrak{ag}}$ is the matrix containing elements $\mathrm{Cov}(\hat{R}(\mathfrak{a}_i), \hat{R}(\mathfrak{g}_j))$. Here I assume that the risk estimate $\hat{R}(\mathfrak{g}_j)$ at the centroid of grid square $\mathfrak{g}_j$ is constant within grid square $\mathfrak{g}_j$, hence $\hat{R}(\mathfrak{g}_j) = \hat{R}(\mathcal{G}_j)$.

Then the conditional distribution of $\hat{\boldsymbol{R}}(\mathfrak{g})|\hat{\boldsymbol{R}}(\mathfrak{a})$ is given by the kriging predictor as follows:

$$\hat{\boldsymbol{R}}(\mathfrak{g}) = \hat{\beta}_0 \boldsymbol{I} + \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_{\mathfrak{a}\mathfrak{g}}^{\top} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_{\mathfrak{a}}^{-1} [\hat{\boldsymbol{R}}(\mathfrak{a}) - \hat{\beta}_0 \mathbf{1}]. \tag{4.2.10}$$

Here $(\hat{\beta}_0, \hat{\boldsymbol{\theta}})$ are estimated by maximum likelihood estimation. Full details of kriging are presented in Section 2.3.5. Then the weights are defined as follows:

$$\omega_{ij} = \frac{e(\mathcal{G}_j)\hat{R}(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} e(\mathcal{G}_k)\hat{R}(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}}, \tag{4.2.11}$$

Multiple imputation algorithm

1. Generate $Y^{(L)}(\mathcal{G}_j)$ for $L = 1, 2, ..., l$ based on $Y(\mathcal{A}_i)$ via multinomial sampling steps, which is outlined in Section 4.2.3. Allison (2000) and Carpenter and Kenward (2008) have been suggested that five imputed datasets ($l = 5$) are sufficient on theoretical grounds. However, $l = 10$ is used in this study in order to reduce sampling variability from the imputation process.

2. For each imputed dataset from the previous step, fit a spatial CAR model proposed by Leroux et al. (2000) to obtain the posterior samples for all model parameters.

3. Combine the results from step 2 in order to make a model inference.

4.2.4 Model

Now I have $[Y(\mathcal{G}_j), e(\mathcal{G}_j), \mathbf{x}(\mathcal{G}_j)]$, the overall model to obtain grid level inference is given by

Grid Level Model

$$
\begin{aligned}
Y(\mathcal{G}_j)|e(\mathcal{G}_j)R(\mathcal{G}_j) &\sim \text{Poisson}[e(\mathcal{G}_j)R(\mathcal{G}_j)] \\
\ln[R(\mathcal{G}_j)] &= \mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j) \\
\phi(\mathcal{G}_j)|\boldsymbol{\phi}(\mathcal{G}_{-j}) &\sim \text{N}\left(\frac{\rho\sum_{k=1}^{m} w_{kj}\phi(\mathcal{G}_k)}{\rho\sum_{k=1}^{m} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{k=1}^{m} w_{kj} + 1 - \rho}\right) \\
\beta_h &\sim \text{N}(0, 100000) \quad \text{for} \quad h = 1, ..., H \\
\tau^2 &\sim \text{Inverse-Gamma}(1, 0.01) \\
\rho &\sim \text{Uniform}(0, 1).
\end{aligned}
\tag{4.2.12}
$$

Here, $R(\mathcal{G}_j)$ denotes disease risk in grid square $\mathcal{G}_j$ and $\boldsymbol{\beta} = [\beta_1, ..., \beta_H]$ is a vector of regression parameters including an intercept term. In addition, $\boldsymbol{\phi}(\mathcal{G}) = [\phi(\mathcal{G}_1), ..., \phi(\mathcal{G}_m)]$ is a vector of random effects, and $\boldsymbol{\phi}(\mathcal{G}_{-j}) = [\phi(\mathcal{G}_1), ..., \phi(\mathcal{G}_{j-1}), \phi(\mathcal{G}_{j+1}), ..., \phi(\mathcal{G}_m)]$. The spatial autocorrelation between these random effects is controlled by an $m \times m$ neighbourhood matrix, $\boldsymbol{W}$, at the grid level. Note that the sharing a common border approach for defining $\boldsymbol{W}$ is not appropriate to use in this model because some grid squares have no neighbours. Therefore, I use the k-nearest neighbour instead as described in Section 2.3.2, and set $w_{kj} = 1$ if grid square $\mathcal{G}_k$ is one of the 4 nearest grid squares to grid square $\mathcal{G}_j$ in term of distance and $w_{kj} = 0$ if not. Here I use the 4 nearest neighbours because it corresponds to the number of neighbours of grid square under the sharing common border assumption which is the most commonly used to create $\boldsymbol{W}$. However, in some cases I may have $w_{kj} = 1$ and $w_{jk} = 0$, for example a region in a remote area may have a "nearest" neighbour quite some distance away, but those neighbours may have several other regions close to themselves. To ensure that the neighbourhod matrix is symmetric, I set $w_{jk} = 1$ in all such cases. Consequently, some grid squares have more than four neighbours and the maximum number of neighbours in this study is eight for both sizes of grid square.

The random effect of grid square $\mathcal{G}_j$, $\phi(\mathcal{G}_j)$ is modelled by a conditional autoregressive model. In this study I am using the model proposed by Leroux et al. (2000) which is described in Section 2.3.4. In this model, ρ controls for the level of spatial autocorrelation. A value of $\rho = 0$ corresponds to a completely spatially smooth model with mean 0 and variance τ^2, while $\rho = 1$ corresponds to the intrinsic model proposed by

Besag et al. (1991). Finally, I use the hyperparameters (0, 100000) in the normal prior for regression parameter, (1, 0.1) in the inverse gamma prior for τ^2 and (0, 1) in the uniform prior for ρ.

4.2.5 Inference

Inference for this model is performed using an MCMC algorithm via a combination of Gibbs sampling and Metropolis-Hastings steps. The parameters are updated in the algorithm as described below.

For each iteration $t = 1, ..., T$ of the MCMC algorithm, conduct the following steps.

- **Update $\boldsymbol{\beta}$.**

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

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{Y}(\mathcal{G}), \boldsymbol{\phi}) &\propto \prod_{j=1}^{m} \text{Poisson}(Y(\mathcal{G}_j)|\boldsymbol{\beta}) \times \prod_{r=1}^{p} \text{N}(\beta_h|0, c) \\
&\propto \prod_{j=1}^{m} \left[\exp(\mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j))\right]^{Y(\mathcal{G}_j)} \times \exp[-\exp(\mathbf{x}^*(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j))] \\
&\times \prod_{r=1}^{p} \exp\left(-\frac{1}{2c}\beta_h^2\right).
\end{aligned}
$$

  Update $\boldsymbol{\beta}$ using the Metropolis-Hastings algorithm, with a proposal $\boldsymbol{\beta}^*$ randomly sampled from the distribution $\boldsymbol{\beta}^* \sim \text{N}\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{V}_\beta\right)$, where $\boldsymbol{\beta}^{(t)}$ is the current value. The acceptance probability of $\boldsymbol{\beta}^*$ is given by $\min\left[1, \frac{f(\boldsymbol{\beta}^*|\boldsymbol{Y}^*(\mathcal{G}), \boldsymbol{\phi}(\mathcal{G}))}{f\left(\boldsymbol{\beta}^{(t)}|\boldsymbol{Y}^*(\mathcal{G}), \boldsymbol{\phi}(\mathcal{G})\right)}\right]$, where $\boldsymbol{\phi}(\mathcal{G}) = (\phi(\mathcal{G}_1), ..., \phi(\mathcal{G}_m))$. The proposal variance $\boldsymbol{V}_\beta$ can be adapted to keep an acceptance rate between 15% and 35% for parameters of high dimension (Roberts et al., 1997).

- **Update $\boldsymbol{\phi} = [\phi(\mathcal{G}_1), ..., \phi(\mathcal{G}_m)]$.**

  Update each $\phi(\mathcal{G}_j)$ in turn for $j = 1, ..., m$ from its full conditional distribution.

$$
\begin{aligned}
f(\phi(\mathcal{G}_j)|Y(\mathcal{G}_j)) &\propto \text{Poisson}(Y(\mathcal{G}_j)|\phi(\mathcal{G}_j)) \\
&\quad \times \text{N}\left(\mu_j = \frac{\rho \sum_{k=1}^{m} w_{kj}\phi(\mathcal{G}_k)}{\rho \sum_{k=1}^{m} w_{kj} + 1 - \rho}, \sigma_j^2 = \frac{\tau^2}{\rho \sum_{k=1}^{m} w_{kj} + 1 - \rho}\right) \\
&\propto \left[\exp(\mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j))\right]^{Y(\mathcal{G}_j)} \times \exp[-\exp(\mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j))] \\
&\quad \times \exp\left[-\frac{1}{2\sigma_j^2}(\phi(\mathcal{G}_j) - \mu_j)^2\right].
\end{aligned}
$$

  Update $\phi(\mathcal{G}_j)$ using the Metropolis-Hastings algorithm, with a proposal $\phi^*(\mathcal{G}_j)$ randomly sampled from the distribution $\phi^*(\mathcal{G}_j) \sim \text{N}\left(\phi^{(t)}(\mathcal{G}_j), v_{\phi(\mathcal{G}_j)}\right)$, where $\phi^{(t)}(\mathcal{G}_j)$ is the current value. The acceptance probability of $\phi^*(\mathcal{G}_j)$ is given by $\min\left[1, \frac{f\left(\phi^*(\mathcal{G}_j)|Y(\mathcal{G}_j), \boldsymbol{\beta}\right)}{f\left(\phi^{(t)}(\mathcal{G}_j)|Y(\mathcal{G}_j), \boldsymbol{\beta}\right)}\right]$. The proposal variance $v_{\phi(\mathcal{G}_j)}$ can be adapted to keep an acceptance rate between 15% and 35%.

- **Update τ^2.**

  The full conditional distribution for τ^2 is as follows:

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

  where $\boldsymbol{Q} = \rho[\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}] + (1-\rho)\boldsymbol{I}$ as described in the CAR model proposed by Leroux et al. (2000) and given in (4.2.12). Gibbs sampling can be used to sample directly from this full conditional distribution.

- **Update ρ.**

  The full conditional distribution for ρ is as follows:

$$\begin{aligned} f(\rho|\boldsymbol{\phi}(\mathcal{G}), \tau^2) &\propto \mathrm{N}(\mathbf{0}, \tau^2\boldsymbol{Q}^{-1}) \times \mathrm{Uniform}(\rho|0,1) \\ &\propto |\boldsymbol{Q}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\phi}(\mathcal{G})^\top \boldsymbol{Q}\boldsymbol{\phi}(\mathcal{G})\right) I_{[\rho\in[0,1]]}. \end{aligned}$$

  Update ρ using the Metropolis-Hastings algorithm, with a proposal ρ^* draw from the distribution $\rho^* \sim \mathrm{N}\left(\rho^{(t)}, v_\rho\right)$ with $0 \leq \rho^* \leq 1$, where $\rho^{(t)}$ is the current value. The acceptance probability of ρ^* is given by $\min\left[1, \frac{f(\rho^*|\boldsymbol{\phi}(\mathcal{G}),\tau^2)}{f(\rho^{(t)}|\boldsymbol{\phi}(\mathcal{G}),\tau^2)}\right]$. The proposal variance v_ρ can be adapted to keep an acceptance rate between 40% and 60% for parameter of low dimension (Roberts et al., 1997).

4.3 Simulation study

4.3.1 Aim

Model (4.2.12) aims to estimate disease risk at the grid level based on data from the areal unit level, and interest lies in estimating disease risk, $R(\mathcal{G}_j)$, and regression parameters, $\boldsymbol{\beta}$. Therefore, a simulation study is conducted to determine how accurately this model can estimate disease risk and regression parameters at the grid level.

4.3.2 General approach

This simulation study consists of four steps. First, I generate observed disease case counts, expected disease case counts and their covariates at the grid level $[Y(\mathcal{G}_j), e(\mathcal{G}_j),$ $\mathrm{x}(\mathcal{G}_j)]$. The method of generating the data will be discussed in Section 4.3.3. Next, I aggregate the grid level data from step one to the areal unit level, because that is what I have for the real data. Then, I fit the model to the grid level data with different scenarios to estimate disease risk and regression parameters at the grid level. Finally, I repeat steps one to three for r simulated datasets and measure how accurate the estimates of disease risk and the regression parameters are. The methods for quantifying these estimates are outlined in Section 4.3.4.

4.3.3 Grid level data generation

To make the simulated data realistic, I base this study on the disease data from the Greater Glasgow and Clyde Health Board which I will use in my application. This data set contains covariates, for example measures of air pollution and poverty. Furthermore, I also know the population at the grid level, $P(\mathcal{G}_j)$, the area of intersection between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$, $a(\mathcal{A}_i \cap \mathcal{G}_j)$, and the neighbourhood matrix at the grid level, $\boldsymbol{W}$. Using this information, I can simulate grid level data, which consist of a disease count $Y(\mathcal{G}_j)$, an expected count $e(\mathcal{G}_j)$ and covariate data $\mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j)$ for each of m grid squares. In this study I create grids using squares with sides of lengths 1,000 and 500 metres, which respectively give 853 and 3,106 grid squares in total. The data are generated from model (4.2.12), with two covariates as follows:

$$\begin{aligned} Y(\mathcal{G}_j) &\sim \text{Poisson}[e(\mathcal{G}_j)R(\mathcal{G}_j)] \qquad j = 1, ..., m \\ \ln[R(\mathcal{G}_j)] &= \beta_1 \mathrm{x}_1(\mathcal{G}_j) + \beta_2 \mathrm{x}_2(\mathcal{G}_j) + \phi(\mathcal{G}_j). \end{aligned} \tag{4.3.1}$$

Here, I want to generate $Y(\mathcal{G}_j)$, which means I must first generate $e(\mathcal{G}_j)$ and $R(\mathcal{G}_j)$. To generate $R(\mathcal{G}_j)$, I have to set the regression parameters (β_1, β_2), and generate covariates and random effects $[\mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j), \phi(\mathcal{G}_j)]$. In general, real data have two types of covariates, those available at the areal unit level (e.g. poverty) and those available at the grid level (e.g. air pollution). Therefore, two covariates are used when generating the data, one where the true grid level values $\mathrm{x}_1(\mathcal{G}_j)$ are unknown, only aggregated areal level values are known $[\mathrm{x}_1(\mathcal{A}_i)]$ and the other where the true grid level values $[\mathrm{x}_2(\mathcal{G}_j)]$ are known. Moreover, after I generate these covariates and fit the model in the simulation study, I can compare the estimates of these regression parameters to see if there are differences in how the model performs in estimating these terms.

When generating the covariates, they are assumed to be normally distributed with mean zero and variance one, $[\mathrm{x}_1(\mathcal{G}_j) \sim \mathrm{N}(0,1), \mathrm{x}_2(\mathcal{G}_j) \sim \mathrm{N}(0,1)]$. The vector of random effects, $\boldsymbol{\phi}(\mathcal{G})$ is generated from a multivariate normal distribution with mean zero and variance $\tau^2\boldsymbol{Q}^{-1}$, $\left[\boldsymbol{\phi}(\mathcal{G}) \sim \mathcal{N}(\boldsymbol{0}, \tau^2\boldsymbol{Q}^{-1})\right]$ which corresponds to the spatial random effects from the conditional autoregressive model proposed by Leroux et al. (2000), with $\boldsymbol{Q} = \rho[\text{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}] + (1-\rho)\boldsymbol{I}$. Here $\boldsymbol{W}$ is the neighbourhood matrix at

Table 4.1: The scenarios used in the simulation study.

Scenario	ρ	τ^2	ψ
1	0.99	0.01	0.01
2	0.99	0.01	0.05
3	0.5	0.01	0.01
4	0.5	0.01	0.05
5	0.99	0.05	0.01
6	0.99	0.05	0.05
7	0.5	0.05	0.01
8	0.5	0.05	0.05

the grid level, and $\boldsymbol{W1}$ is a vector containing the number of neighbours for each grid square. In the simulation study ρ and τ^2 are varied in the simulation design, while β_1 and β_2 are fixed at 0.1 ($\beta_1 = \beta_2 = 0.1$).

Next, I want to estimate $e(\mathcal{G}_j)$, the expected disease count for grid square $\mathcal{G}_j$. Here I use the adjusted population in each grid square, $P(\mathcal{G}_j)$, and assume that $e(\mathcal{G}_j) = \psi P(\mathcal{G}_j)$, where ψ is the proportion of the population who have the disease event. The value of ψ was varied in the simulation study as well as ρ and τ^2. I simulate data under different scenarios as can be seen in Table 4.1.

I simulate data under different scenarios with different level of prevalence of the disease and different level of variation in disease risk in order to explore the effects when increasing the prevalence and variation in disease risks across the region has on the estimation. the variety of scenarios which might occur in real data. Here, $\rho = 0.99$ and 0.5 respectively represent strong and moderate spatial dependence. I use $\tau^2 = 0.01$ and 0.05 to compare the accuracy of the estimates under different levels of variation. I compare $\psi = 0.01$ and 0.05 to test the model under different disease frequencies. note that, the minimal spatial autocorrelation is rarely found in real situation (disease risk) and the results from Chapter 3 indicate that the spatial autocorrelation is strong ($rho = 0.74$) therefore the minimal autocorrelation is not included in the simulation studies. Now all the grid level data $[Y(\mathcal{G}_j), e(\mathcal{G}_j), R(\mathcal{G}_j), \mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j), \phi(\mathcal{G}_j)]$ needed for this simulation study are set.

4.3.4 Data aggregation

In section 4.3.3, I generated grid level data, and now I must aggregate them to the areal unit level to create the data appropriate for the proposed method. The disease cases at the areal unit level $e(\mathcal{A}_i)$ can be estimated as

$$e(\mathcal{A}_i) = \sum_{j=1}^{m} e(\mathcal{A}_i \cap \mathcal{G}_j) = \sum_{j=1}^{m} \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{n} a(\mathcal{A}_k \cap \mathcal{G}_j)} e(\mathcal{G}_j), \tag{4.3.2}$$

so that the expected cases for each grid square are allocated to regions based on the proportion of the grid square that lies in that region. This is a weighted average based on the area of intersection between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$. It is straightforward to show that $\sum_{j=1}^{m} e(\mathcal{G}_j) = \sum_{i=1}^{n} e(\mathcal{A}_i)$. Next I must quantify the disease count in region $\mathcal{A}_i$, $Y(\mathcal{A}_i)$. This follows a very similar argument to $e(\mathcal{A}_i)$, except that the disease counts must be a non-negative integer. Therefore I set

$$Y(\mathcal{A}_i) = \left\lfloor \sum_{j=1}^{m} \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{n} a(\mathcal{A}_k \cap \mathcal{G}_j)} Y(\mathcal{G}_j) \right\rceil, \tag{4.3.3}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. Note that, $\sum_{i=1}^{n} Y(\mathcal{A}_i)$ may not exactly equal $\sum_{j=1}^{m} Y(\mathcal{G}_j)$ due to this rounding but they will be similar. Finally, I have to aggregate the covariate, $\mathrm{x}_1(\mathcal{G}_j)$ to the areal unit level. Here I have a continuous covariate, and I can therefore aggregate it as follows:

$$\mathrm{x}_1(\mathcal{A}_i) = \sum_{j=1}^{m} \frac{P(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{m} P(\mathcal{A}_i \cap \mathcal{G}_k)} \mathrm{x}_1(\mathcal{G}_j). \tag{4.3.4}$$

This is a weighted average based on the population in the intersection area between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$, so that grid squares with larger populations have greater weight. However, $P(\mathcal{A}_i \cap \mathcal{G}_j)$ is unknown, so I estimate it based on the assumption of a common population density across each grid square. Therefore, under this assumption, I estimate $P(\mathcal{A}_i \cap \mathcal{G}_j) = P(\mathcal{G}_j) a(\mathcal{A}_i \cap \mathcal{G}_j) / \sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)$, the population in grid square $\mathcal{G}_j$ multiplied by the proportion of region $\mathcal{A}_i$ in grid square $\mathcal{G}_j$. I therefore have

$$\mathrm{x}_1(\mathcal{A}_i) = \sum_{j=1}^{m} \frac{P(\mathcal{G}_j) \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} P(\mathcal{G}_k) \frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}} \mathrm{x}_1(\mathcal{G}_j). \tag{4.3.5}$$

I assume $\mathrm{x}_2(\mathcal{G}_j)$ is known when fitting the grid level model, hence no data aggregation is needed for $\mathrm{x}_2(\mathcal{G}_j)$.

I have now generated the aggregated data $[Y(\mathcal{A}_i), e(\mathcal{A}_i), \mathrm{x}_1(\mathcal{A}_i)]$. In my application, I would only have access to this data plus the grid level covariate $\mathrm{x}_2(\mathcal{G}_j)$, and the goal would be to estimate $Y(\mathcal{G}_j), R(\mathcal{G}_j), \beta_1$ and β_2. Here, I fit the model to the aggregated data to test the accuracy of this estimation.

4.3.5 Fitting the model

The next step is to fit model (4.2.12) under two different test scenarios. In scenario one (Model 1), I fit the model to the true grid level data $[Y(\mathcal{G}_j), e(\mathcal{G}_j), \mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j)]$. In scenario two, I have the aggregated data $[Y(\mathcal{A}_i), e(\mathcal{A}_i), \mathrm{x}_1(\mathcal{A}_i)]$, and $\mathrm{x}_2(\mathcal{G}_j)$, as would be the case in the real study. In this case I need to disaggregate them to the grid level $[\tilde{Y}(\mathcal{G}_j), \tilde{e}(\mathcal{G}_j), \tilde{\mathrm{x}}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j)]$. There are however, two different methods to estimate $Y(\mathcal{G}_j)$ as described in Section 4.2.3. Model 2 uses (4.2.9) and Model 3 uses (4.2.11) to estimate $Y(\mathcal{G}_j)$ via multinomial steps. In summary, the three models are described as follows

- Model 1 - fit the model to the true grid level data.
- Model 2 - fit the model to the disaggregated data at grid level with $\hat{R}(\mathcal{G}_j) = 1$ for all j.
- Model 3 - fit the model to the disaggregated data at grid level with disease risk, $\hat{R}(\mathcal{G}_j)$ estimated via kriging.

Model 1 should perform the best of the three models since it is fitted to the true grid level data. It therefore acts as a reference model to compare the performance of Models 2 and 3. I repeat steps one to three outlined in Section 4.3.2 to generate r datasets, and fit each model to each dataset.

In this study, I generate 100 simulated datasets ($r = 100$), and for each of them I generate 10 imputed datasets. Then fit the models to each imputed datasets, and the final step is to combine the results for all imputed datasets to make a model inference. The

aim is to estimate the regression parameters and disease risk $[\hat{\beta}_{hl}, \hat{R}_l(\mathcal{G}_j)]$ for $h = 1, 2$ and $l = 1, ..., r$, from the posterior median of the parameters estimates from the MCMC iterations of the lth dataset. I also wish to measure variation and therefore observe the upper and lower limits of the 95% credible interval. These are the 2.5th and 97.5th percentile from the MCMC iterations of the lth dataset. Here inference will be based on 200,000 MCMC samples with 50,000 burn-in iterations, with the remaining 150,000 samples thinned by a factor of 15. Therefore, a total of 100,000 samples are used for model inference, with 10,000 samples for each imputed dataset. I compare the results from Model 1, Model 2 and Model 3 to see how accurate each model is.

4.3.6 Summarising the results

To measure how accurately each model estimates disease risk and regression parameters at the grid level, I use four metrics; bias, root mean square error (RMSE), credible interval (CI) coverage, and average width.

Bias

Bias is used to quantify the average difference between the estimated values and the true value. These individual differences are also called residuals. The bias of the regression parameters, (β_1, β_2), and disease risk at the grid level, $R(\mathcal{G})$, are given by

$$\begin{aligned} \text{Bias}(\beta_h) &= \frac{1}{r}\sum_{l=1}^{r}\left(\hat{\beta}_{hl} - \beta_h\right) \\ \text{Bias}(R(\mathcal{G})) &= \frac{1}{rm}\sum_{j=1}^{m}\sum_{l=1}^{r}\left[\hat{R}_l(\mathcal{G}_j) - R_l(\mathcal{G}_j)\right], \end{aligned} \tag{4.3.6}$$

where $(\hat{\beta}_{hl}, \hat{R}_l(\mathcal{G}_j))$ represent estimates of the true values $(\beta_h, R_l(\mathcal{G}_j))$ for the lth simulated data set. Bias measures whether the average of the estimates is greater or less than the true value. It does not tell us about the precision of the estimates. For example, suppose I generate two data sets and fit the model, and I obtain residuals of -10 and 10: the bias in this case is equal to zero which I expect from a good model but these estimates are not precise. Therefore, I need to consider other measurements simultaneously to quantify the precision of the estimates.

Root mean square error (RMSE)

RMSE measures the variation in the estimated values around the true value by taking the mean of the square of the residuals. Smaller RMSE values indicate more precise estimation. The RMSE of (β_1, β_2) and $R(\mathcal{G})$ are given by

$$\begin{aligned} \text{RMSE}(\beta_h) &= \sqrt{\frac{1}{r}\sum_{l=1}^{r}\left(\hat{\beta}_{hl} - \beta_h\right)^2} \\ \text{RMSE}(R(\mathcal{G})) &= \sqrt{\frac{1}{rm}\sum_{j=1}^{m}\sum_{l=1}^{r}\left[\hat{R}_l(\mathcal{G}_j) - R_i(\mathcal{G}_j)\right]^2}. \end{aligned} \tag{4.3.7}$$

CI coverage

I am also interested in measuring how accurately a model accounts for uncertainty. I can measure this by looking at how often the true value lies within the 95% credible interval from the model. For data set l, I can denote the upper and lower bounds of the 95% credible interval respectively by $U_l(\cdot)$ and $L_l(\cdot)$. Next, I compute the proportion of these credible interval which contain the true value for each parameter. If the proportion is close to 0.95 then the model is appropriately quantifying uncertainty. This can be written as,

$$\begin{aligned} \text{CI coverage}(\beta_h) &= p\left(L_l(\hat{\beta}_h) \leq \beta_h \leq U_l(\hat{\beta}_h)\right) \\ \text{CI coverage}(R(\mathcal{G})) &= p\left(L_l(\hat{R}(\mathcal{G}_j)) \leq R(\mathcal{G}_j) \leq U_l(\hat{R}(\mathcal{G}_j))\right), \\ &\qquad \text{for} \quad l = 1, ..., r \quad \text{and} \quad j = 1, ..., m, \end{aligned} \tag{4.3.8}$$

where $p(\cdot)$ denotes the proportion of the 95% credible intervals that contain the true value.

Average width

The average width of a credible interval can be used to compare models in terms of the precision of interval estimation. If I have two models with similar CI coverages, the model with a narrower average width provides more precise estimation. The average

width of (β_1, β_2) and $R(\mathcal{G})$ are given by

$$\begin{aligned} \text{Average width}(\beta_h) &= \frac{1}{r}\sum_{l=1}^{r}\left[U_l(\hat{\beta}_h) - L_l(\hat{\beta}_h)\right] \\ \text{Average width}(R(\mathcal{G})) &= \frac{1}{rm}\sum_{j=1}^{m}\sum_{l=1}^{r}\left[U_l(\hat{R}(\mathcal{G}_j)) - L_l(\hat{R}(\mathcal{G}_j))\right]. \end{aligned} \tag{4.3.9}$$

4.3.7 Simulation results

I created grid square with sides of lengths 1,000 and 500 metres and simulated one hundred data sets for each model and for each of these simulations I recorded bias, RMSE, CI coverage, and average width of CI. The simulation results for the grid squares with 1,000 metres length are shown in Tables 4.2, 4.3 and 4.4, while for the grid squares with 500 metres length are shown in Tables 4.5, 4.6 and 4.7.

Table 4.2: Results from the simulation study for the regression parameter $\beta_1 = 0.1$ with the estimated covariate at the grid level $\mathrm{x}_1(\mathcal{G}_j)$.

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	-0.0012	-0.0019	0.0074
	2	0.0009	0.0021	0.0119
	3	0.0011	-0.0046	0.0048
	4	-0.0003	0.0003	0.0067
	5	0.0001	0.0020	0.0178
	6	-0.0021	0.0039	0.0228
	7	0.00022	-0.0014	0.0088
	8	-0.0006	-0.0007	0.0095
RMSE	1	0.0110	0.0213	0.0241
	2	0.0057	0.0099	0.0162
	3	0.0092	0.0187	0.0223
	4	0.0056	0.0120	0.0169
	5	0.0129	0.0180	0.0277
	6	0.0068	0.0177	0.0318
	7	0.0144	0.0213	0.0247
	8	0.0083	0.0176	0.0229
CI coverage	1	0.94	0.98	0.98
	2	0.94	0.98	0.86
	3	0.94	0.98	0.98
	4	0.92	0.96	0.86
	5	0.94	1.00	1.00
	6	0.99	0.94	0.72
	7	0.90	0.96	0.96
	8	0.96	0.94	0.82
Average width	1	0.0401	0.0958	0.0977
	2	0.0211	0.0527	0.0553
	3	0.0398	0.0917	0.0942
	4	0.0219	0.0501	0.0515
	5	0.0468	0.1119	0.1188
	6	0.0301	0.0695	0.0761
	7	0.0492	0.1032	0.1096
	8	0.0338	0.0647	0.0698

Table 4.3: Results from the simulation study for the regression parameter $\beta_2 = 0.1$ with the true covariate at the grid level $x_2(\mathcal{G}_j)$.

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	0.0003	-0.0682	-0.0657
	2	-0.0003	-0.0711	-0.0685
	3	0.0004	-0.0662	-0.0628
	4	0.0005	-0.0706	-0.0687
	5	-0.0016	-0.0719	-0.0676
	6	0.0002	-0.0760	-0.0719
	7	0.0030	-0.0689	-0.0659
	8	0.0015	-0.0754	-0.0729
RMSE	1	0.0117	0.0685	0.0661
	2	0.0058	0.0711	0.0686
	3	0.0101	0.0665	0.0631
	4	0.0047	0.0707	0.0688
	5	0.0104	0.0721	0.0679
	6	0.0064	0.0762	0.0720
	7	0.0131	0.0693	0.0663
	8	0.0092	0.0755	0.0731
CI coverage	1	0.96	0	0
	2	0.92	0	0
	3	0.98	0	0
	4	1.00	0	0
	5	0.98	0	0
	6	0.97	0	0
	7	0.94	0	0
	8	0.92	0	0
Average width	1	0.0406	0.0501	0.0496
	2	0.0209	0.0245	0.0249
	3	0.0402	0.0494	0.0495
	4	0.0220	0.0243	0.0248
	5	0.0471	0.0522	0.0543
	6	0.0301	0.0280	0.0295
	7	0.0486	0.0509	0.0528
	8	0.0339	0.0269	0.0289

Table 4.4: Results from the simulation study for the disease risk at the grid level $R(\mathcal{G}_j)$.

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	-0.0045	-0.0068	-0.0069
	2	0.0000	-0.0018	-0.0019
	3	0.0012	-0.0008	-0.0012
	4	0.0004	-0.0017	-0.0019
	5	-0.0021	-0.0109	-0.0123
	6	-0.0015	-0.0081	-0.0081
	7	0.0014	-0.0037	-0.0054
	8	0.0025	-0.0015	-0.0023
RMSE	1	0.0802	0.1438	0.1423
	2	0.0649	0.1394	0.1376
	3	0.0713	0.1361	0.1345
	4	0.0633	0.1371	0.1362
	5	0.1506	0.2023	0.1976
	6	0.1236	0.1948	0.1899
	7	0.1437	0.1948	0.1926
	8	0.1246	0.1920	0.1900
CI coverage	1	0.95	0.67	0.72
	2	0.95	0.57	0.62
	3	0.94	0.57	0.63
	4	0.95	0.51	0.55
	5	0.95	0.74	0.82
	6	0.95	0.64	0.71
	7	0.94	0.62	0.71
	8	0.95	0.59	0.65
Average width	1	0.3173	0.2728	0.2995
	2	0.2479	0.2182	0.2401
	3	0.2695	0.2154	0.2403
	4	0.2459	0.1871	0.2031
	5	0.5634	0.4439	0.5134
	6	0.4443	0.3397	0.3840
	7	0.5462	0.3406	0.4050
	8	0.4602	0.3132	0.3468

The results for all metrics and models for the grid squares with sides of length 1,000 metres are presented in Tables 4.2, 4.3 and 4.4. The results show consistent messages across all scenarios, which are described below. Overall, Model 1 performs the best for all regression parameters and disease risk [β_1, β_2 and $R(\mathcal{G}_j)$]. All biases are close to zero, and Model 1 produces the smallest RMSE values across the 3 models which indicates more precise estimation. In addition, the CI coverages are close to 0.95 so they suggest the model is appropriate to quantify uncertainty. Collectively these results are not surprising, as Model 1 is fitted to the true grid level data which is thus expected to be the best model. The magnitude of the differences differs depending on the metric, but roughly the RMSE for β_1 and $R(\mathcal{G}_j)$ doubles from Model 1 compared to Models 2 and 3, whereas for β_2 the increase is between a factor of 5 and 15.

In order to compare the results from the proposed models (Model 2 and Model 3), I consider all metrics for the regression parameters and disease risk. I found that both models produce close to unbiased estimates of β_1 which is the parameter relating to the estimated covariate $\mathrm{x}_1(\mathcal{G}_j)$, while they produce biased estimates of β_2 which is the parameter relating to the true known covariate $\mathrm{x}_2(\mathcal{G}_j)$. This is resulting in zero percentage of CI coverages for all scenarios. This result is initially surprising because the true covariate data would be expected to perform better than the estimated covariate. However, the reason for this is because the disease counts at the grid level $Y(\mathcal{G}_j)$ are unknown but $\mathrm{x}_2(\mathcal{G}_j)$ is known, therefore we need to estimate grid level disease counts based on the areal unit disease counts but we do not do this process to $\mathrm{x}_2(\mathcal{G}_j)$. Hence, the level of the relationship between $Y(\mathcal{G}_j)$ and $\mathrm{x}_2(\mathcal{G}_j)$ might be changed. Unlike the unknown covariate at the grid level $\mathrm{x}_1(\mathcal{G}_j)$, we do the transformation process based on areal data for both $Y(\mathcal{G}_j)$ and $\mathrm{x}_1(\mathcal{G}_j)$and therefore the relationship between these two variables is still maintained at the similar level as the areal data.

Model 3 performs better than Model 2 in term of RMSE for β_2 as shown in Table 4.3, since Model 3 has RMSE values slightly smaller than Model 2 for all scenarios. This is because in Model 3 the disease risk at the grid level $R(\mathcal{G}_j)$ is estimated via kriging, and that make the estimated numbers of disease cases $Y(\mathcal{G}_j)$ closer to the true grid level data than when one assumes that $\hat{R}(\mathcal{G}_j) = 1$ (Model 2). Thus as $\mathrm{x}_2(\mathcal{G}_j)$ is the

true grid level values one has smaller error in the parameter estimate than if smoothed disease cases (via $\hat{R}(\mathcal{G}_j) = 1$) are used in the regression. On the other hand, Model 3 performs worse than Model 2 for β_1 as shown in Table 4.2. This is because $Y(\mathcal{G}_j)$ is disaggregated from the areal level data $Y(\mathcal{A}_i)$ and is hence spatially smoother than the true grid level disease counts, which thus aligns better with the spatial smoothing induced by assuming $\hat{R}(\mathcal{G}_j) = 1$ in Model 2.

Furthermore, Model 3 has better estimates of disease risk than Model 2 as measured by RMSE, even though they both produce close to unbiased estimates. This again is likely to be because the Kriging of the risk in Model 3 results in grid level disease counts (via a multinomial sampling step) that are closer to the true values than naively assuming that $\hat{R}(\mathcal{G}_j) = 1$. Additionally this results in coverages that are slightly higher than for Model 2.

Table 4.5: Results from the simulation study for the regression parameter $\beta_1 = 0.1$ with the estimated covariate at the grid level $x_1(\mathcal{G}_j)$ (grid size = 500 m).

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	-0.0018	0.0006	0.0020
	2	-0.0003	0.0052	0.0070
	3	-0.0038	-0.0040	-0.0030
	4	-0.0011	0.0021	0.0037
	5	0.0001	-0.0019	-0.0011
	6	0.0009	0.0022	0.0075
	7	-0.0013	-0.0010	0.0026
	8	0.0009	-0.0005	0.0027
RMSE	1	0.0046	0.0192	0.0193
	2	0.0046	0.0167	0.0174
	3	0.0089	0.0240	0.0252
	4	0.0045	0.0122	0.0134
	5	0.0102	0.0252	0.0257
	6	0.0046	0.0194	0.0223
	7	0.0105	0.0261	0.0282
	8	0.0055	0.0205	0.0209
CI coverage	1	1.00	1.00	1.00
	2	0.93	1.00	0.97
	3	1.00	1.00	1.00
	4	1.00	1.00	1.00
	5	0.97	1.00	1.00
	6	1.00	0.94	0.91
	7	0.94	1.00	1.00
	8	0.98	0.90	0.91
Average width	1	0.0380	0.1368	0.1374
	2	0.0180	0.0717	0.0730
	3	0.0377	0.1368	0.1349
	4	0.0181	0.0652	0.0658
	5	0.0400	0.1541	0.1597
	6	0.0214	0.0842	0.0863
	7	0.0402	0.1357	0.1378
	8	0.0229	0.0753	0.0771

Table 4.6: Results from the simulation study for the regression parameter $\beta_2 = 0.1$ with the true covariate at the grid level $x_2(\mathcal{G}_j)$ (grid size = 500 m).

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	0.0001	-0.0867	-0.0867
	2	-0.0007	-0.0894	-0.0889
	3	-0.0018	-0.0868	-0.0861
	4	0.0000	-0.0878	-0.0879
	5	-0.0013	-0.0898	-0.0880
	6	0.0005	-0.0910	-0.0908
	7	0.0020	-0.0866	-0.0867
	8	-0.0010	-0.0896	-0.0897
RMSE	1	0.0101	0.0868	0.0868
	2	0.0043	0.0895	0.0889
	3	0.0096	0.0869	0.0862
	4	0.0053	0.0879	0.0879
	5	0.0098	0.0899	0.0881
	6	0.0054	0.0910	0.0908
	7	0.0106	0.0867	0.0869
	8	0.0055	0.0897	0.0898
CI coverage	1	0.96	0	0
	2	1.00	0	0
	3	0.98	0	0
	4	0.95	0	0
	5	1.00	0	0
	6	0.96	0	0
	7	0.93	0	0
	8	0.94	0	0
Average width	1	0.0380	0.0479	0.0501
	2	0.0182	0.0233	0.0228
	3	0.0378	0.0470	0.0478
	4	0.0182	0.0221	0.0228
	5	0.0403	0.0505	0.0492
	6	0.0214	0.0233	0.0237
	7	0.0405	0.0491	0.0507
	8	0.0230	0.0227	0.0242

Table 4.7: Results from the simulation study for the disease risk at the grid level $R(\mathcal{G}_j)$ (grid size = 500 m).

Metric	**Scenario**	**Model 1** (True data)	**Model 2** ($\hat{R}(\mathcal{G}_j) = 1$)	**Model 3**($\hat{R}(\mathcal{G}_j)$ via kriging)
Bias	1	-0.0025	-0.0053	-0.0055
	2	0.0007	-0.0011	-0.0006
	3	0.0027	0.0016	0.0016
	4	0.0005	-0.0002	-0.0004
	5	-0.0022	-0.0112	-0.0120
	6	0.0030	-0.0043	-0.0051
	7	-0.0001	-0.0027	-0.0033
	8	-0.0008	-0.0039	-0.0041
RMSE	1	0.0923	0.1635	0.1630
	2	0.0795	0.1587	0.1582
	3	0.0737	0.1515	0.1511
	4	0.0693	0.1517	0.1517
	5	0.1840	0.2399	0.2380
	6	0.1591	0.2322	0.2287
	7	0.1567	0.2110	0.2109
	8	0.1410	0.2096	0.2095
CI coverage	1	0.92	0.61	0.64
	2	0.95	0.57	0.59
	3	0.94	0.57	0.58
	4	0.95	0.46	0.47
	5	0.95	0.72	0.77
	6	0.95	0.64	0.67
	7	0.94	0.53	0.56
	8	0.95	0.53	0.55
Average width	1	0.3241	0.2772	0.2954
	2	0.2987	0.2510	0.2628
	3	0.2880	0.2363	0.2384
	4	0.2744	0.1863	0.1926
	5	0.6639	0.5079	0.5515
	6	0.5589	0.4067	0.4358
	7	0.5825	0.3010	0.3276
	8	0.5363	0.3006	0.3136

To assess the robustness of these results to grid square size, I re-ran the simulation study with grid squares with sides of lengths 500 metres. The results from all metrics and models are presented in Tables 4.5, 4.6 and 4.7. The general pattern of the results is similar to the 1,000 metres grid square results, as the grid square size has not changed the main findings outlined above. However, the results from the models fitted to data with grid square size 500 metres are worse than grid square size 1,000 metres for all metrics. This is true for all models (even Model 1), with for example the RMSEs in disease risk for scenario 1 being: Model 1 - 0.0802 (1,000m) vs 0.0923 (500m); Model 2 - 0.1438 (1,000m) vs 0.1635 (500m); Model 3 -0.1423 (1,000m) vs 0.1630 (500m). This is because the number of grid squares with size 1,000 metres is fewer than the number of grid squares with size 500 metres, which means that I have to estimate more grid level disease counts and risk estimates for the latter, resulting in less accurate estimates.

4.4 Application to real data

In order to illustrate the proposed methodology in this chapter, these models from Section 4.3.5 are applied to data on respiratory disease in the Greater Glasgow and Clyde Health Board for January 2015 to December 2016. Note that only Models 2 and 3 are used in the application, since Model 1 is fitted to the true grid level data which I do not have for the real data.

4.4.1 Data description

The study region is the Greater Glasgow and Clyde Health Board area which is the same region to the previous chapter. The health board is split up into $n = 257$ administrative units called intermediate zones (IZ), containing populations between 1,321 and 9,008 people with a median population of 4,306 (Scottish Government, 2019). They are the same units used in the simulation study. The disease data, $\boldsymbol{Y}(\mathcal{A}) = [Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$, are obtained from the Scottish Statistics website https://statistics.gov.scot, where $Y(\mathcal{A}_i)$ denotes the number of hospital admissions with respiratory disease in region i. The expected disease cases, $\boldsymbol{e}(\mathcal{A}) = [e(\mathcal{A}_1), ..., e(\mathcal{A}_n)]$, are the expected hospital admission numbers of respiratory disease for each region and is computed by indirect standardisation. Figure 4.5 presents the Standardised

Incidence Ratio (SIR) for the respiratory hospital admissions, which is the ratio of the observed disease data to the expected disease cases in each region [SIR $= Y(\mathcal{A}_i)/e(\mathcal{A}_i)$]. It shows that the regions of high risk are located on the east of the city centre and also the south of the Clyde river. In contrast, the regions of low risk are located on the area of the West End (just the north of the Clyde river) and also on the far south of the city centre.

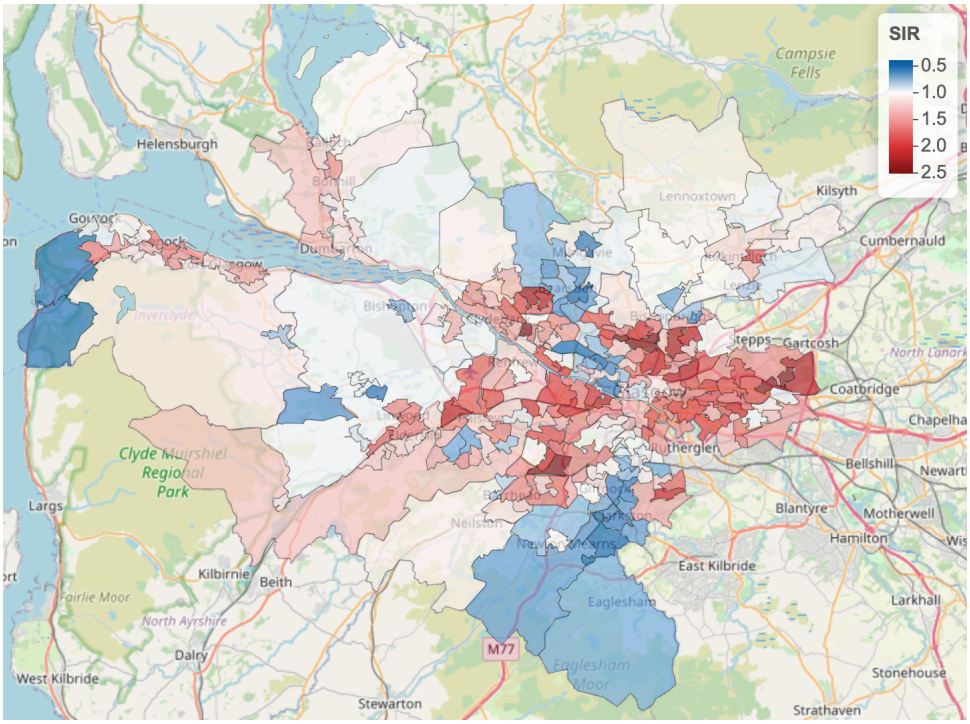


Figure 4.5: The standardised incidence ratio for respiratory disease hospitalisation in Greater Glasgow.

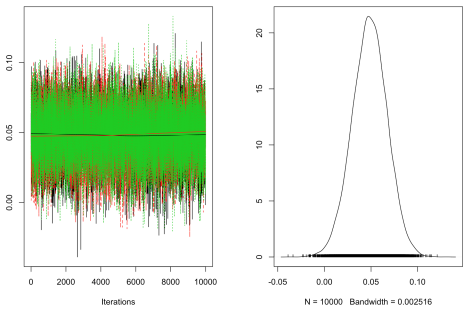
4.4.2 Results

In order to see the performance of the proposed models to the real data, Models 2 and 3 mentioned in Section 4.3.5 are fitted to the respiratory disease data in the Greater Glasgow and Clyde Health Board. Here the grid level data are estimated at the grid squares of sizes of 1,000 and 500 metres via multiple imputation approach, with ten imputed datasets. Markov Chain Monte Carlo inference is used to obtain the results, and the models are run three times to generate MCMC samples from three independent Markov chains. Each chain is run for 200,000 samples, with 50,000 burn-in period and the remaining 150,000 are thinned by a factor of 15. This leaves 300,000 samples for

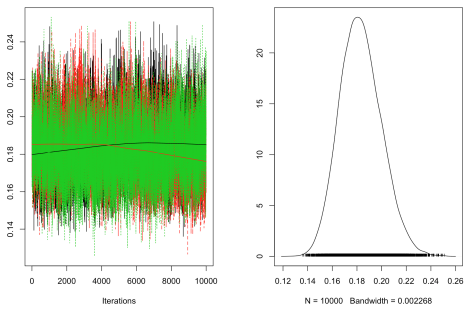
the model inference overall, with 10,000 for each chain and 30,000 for each imputed dataset.

Convergence diagnostic

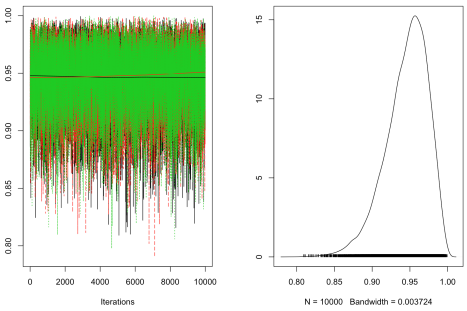
The convergence of the posterior distribution is diagnosed via the method proposed by Gelman and Rubin (1992), and trace plots assessment. Figures 4.6 to 4.9 present trace plots of each model parameter from the proposed models (Models 2 and 3) with the grid square of sizes of 1,000 and 500 meters from one imputed dataset. The figures show that all the chains appear to have converged as there is no clear pattern in the plots. In addition, the trace plots of each parameter for the other nine datasets are very similar to Figures 4.6 to 4.9, therefore they are not shown. An additional check is the Gelman-Rubin (Gelman and Rubin, 1992) diagnostic, which relies on multiple chains and they suggest that less than 1.1 of the value indicates good mixing of the chain. The Gelman-Rubin statistics for all selected parameters are less than 1.1, which indicate that the posterior samples are well mixed.



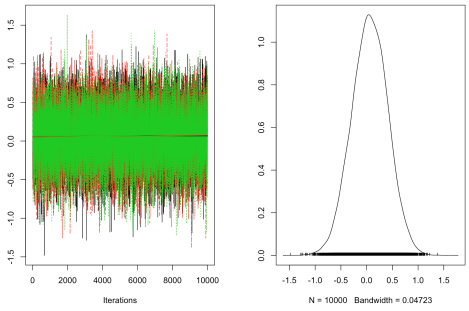
(a) β



(b) τ^2

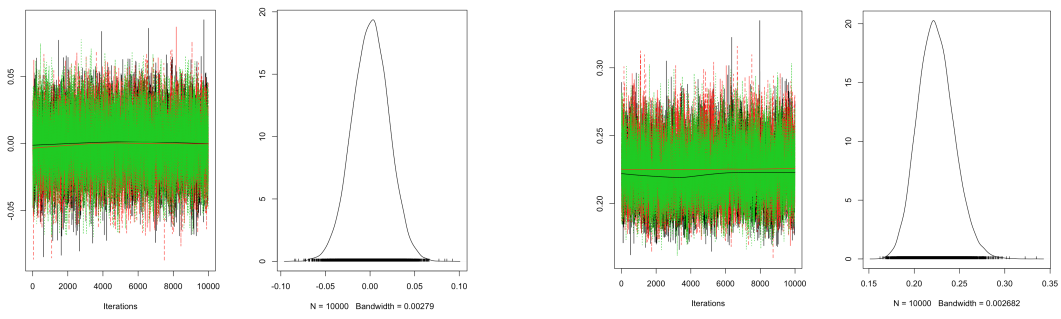


(c) ρ

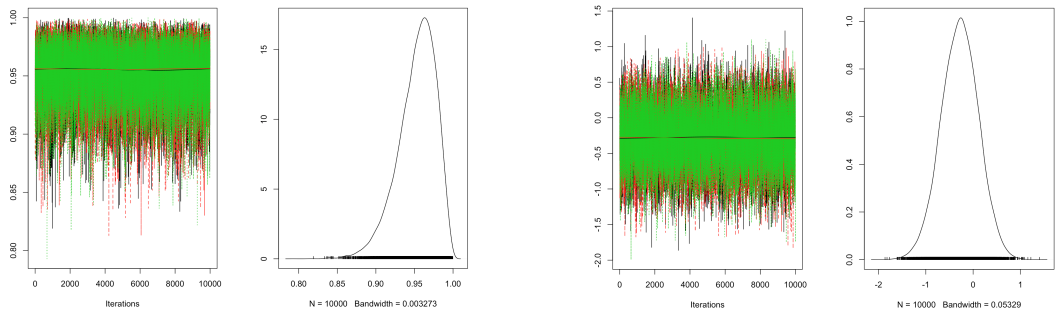


(d) $\phi(\mathcal{G}_1)$

Figure 4.6: Traceplots of MCMC samples for each parameter from Model 2 (grid of size 1,000m).

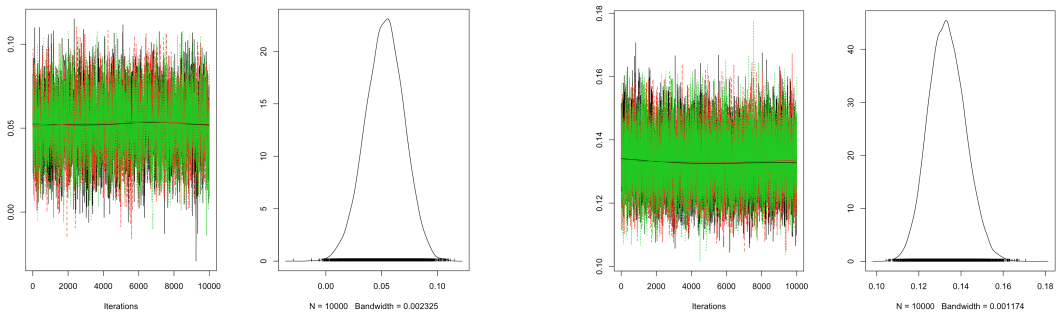


(a) β (b) τ^2

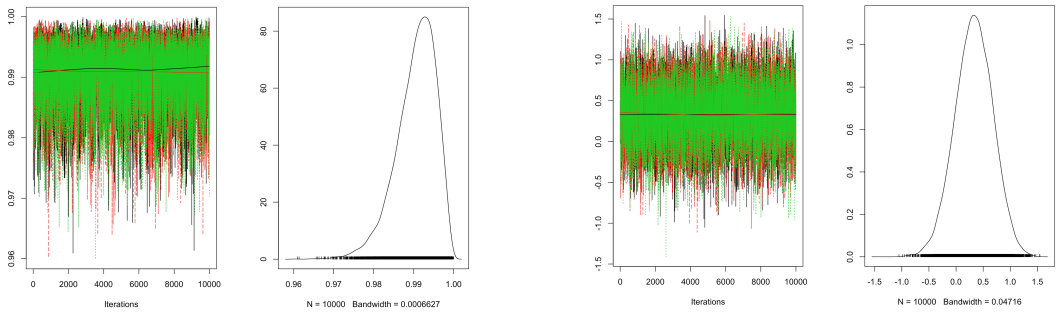


(c) ρ (d) $\phi(\mathcal{G}_1)$

Figure 4.7: Traceplots of MCMC samples for each parameter from Model 3 (grid of size 1,000m).

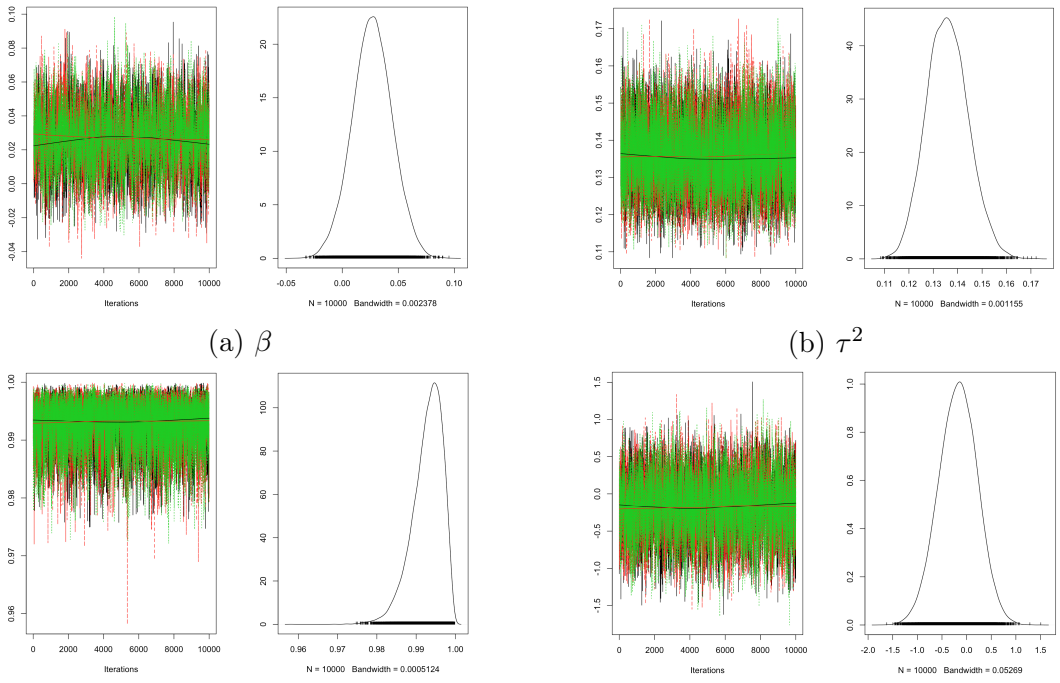


(a) β (b) τ^2



(c) ρ (d) $\phi(\mathcal{G}_1)$

Figure 4.8: Traceplots of MCMC samples for each parameter from Model 2 (grid of size 500m).



(a) β (b) τ^2

(c) ρ (d) $\phi(\mathcal{G}_1)$

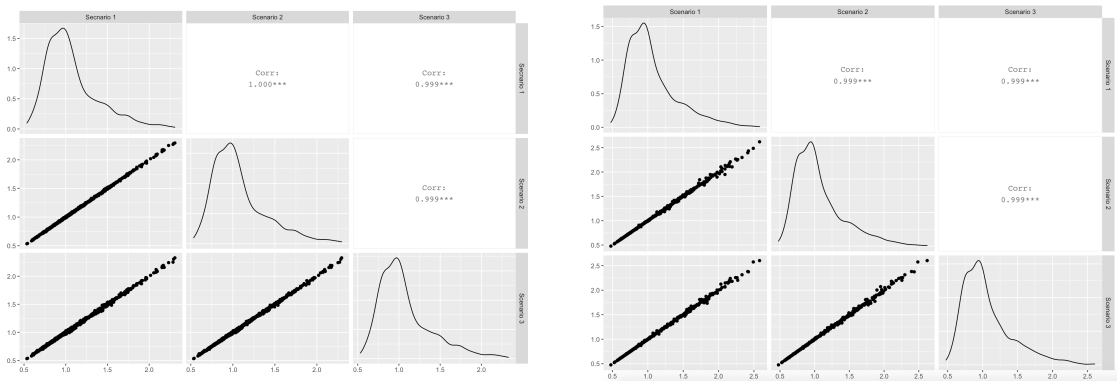
Figure 4.9: Traceplots of MCMC samples for each parameter from Model 3 (grid of size 500m).

Sensitivity analysis

Sensitivity analysis is conducted in order to examine if changes in the model settings, resulting in changes in posterior inferences. There are three hyperpriors for the variance of random effects τ^2 from 4.2.12 being used in the model settings, which are the same as the previous chapter.

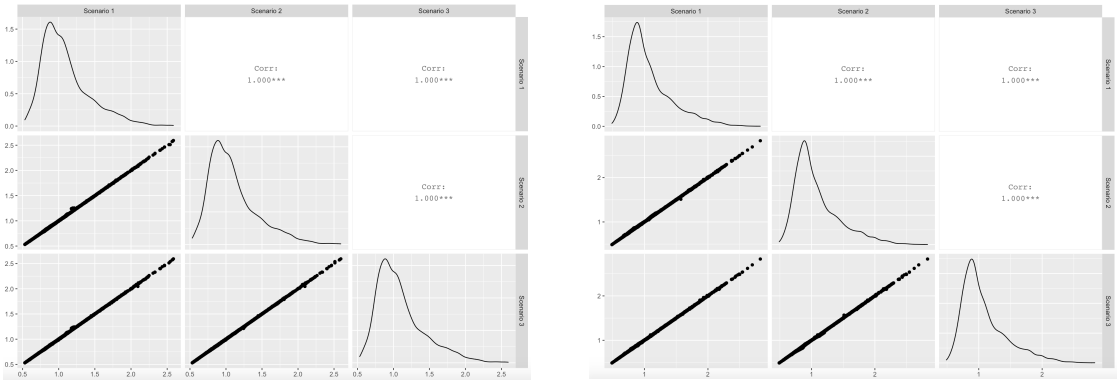
1. Scenario 1 - $\tau^2 \sim$ Inverse-Gamma$(1, 0.01)$.
2. Scenario 2 - $\tau^2 \sim$ Inverse-Gamma$(0.01, 0.01)$.
3. Scenario 3 - $\tau^2 \sim$ Inverse-Gamma$(0.05, 0.0005)$.

Figures 4.10 and 4.11 present the relationship plots of the estimated risks between the choices of hyperpriors for grid squares of sizes of 1,000 and 500 metres respectively. The figures show that the estimated risks among the scenarios lie on the straight line. It means that the choice of hyperpriors does not affect the posterior inferences. Therefore only one hyperprior setting is used for model inference, hence Scenario 1 is randomly selected.



(a) Model2 (b) Model3

Figure 4.10: The estimated risks scatter plots of scenarios 1 - 3 for Models 2 and 3 (grid of size 1,000m).

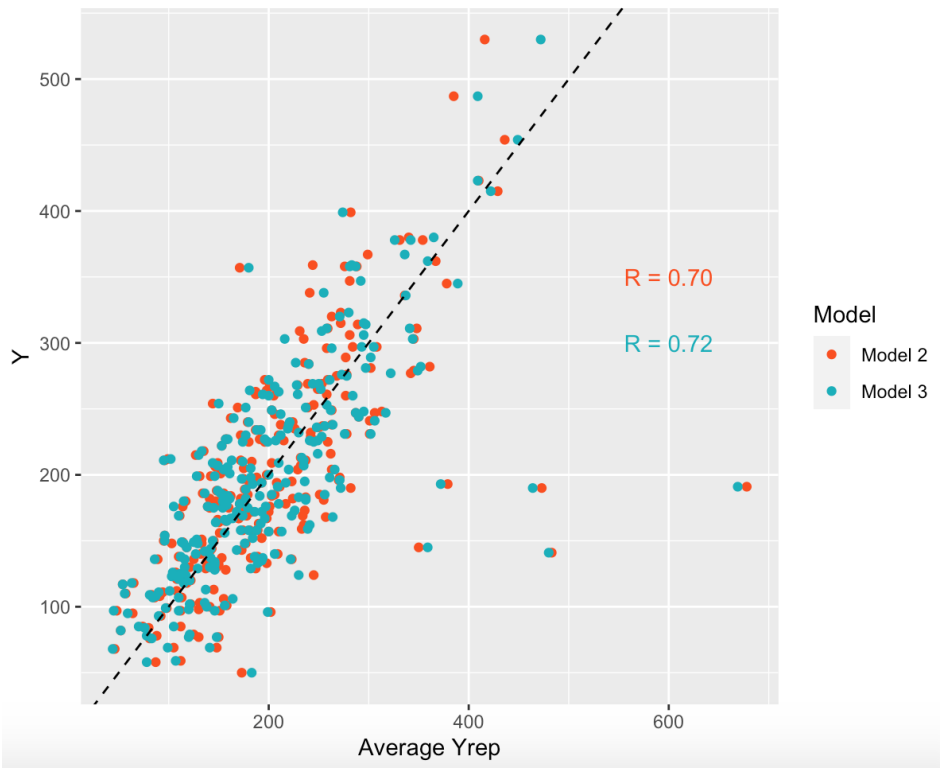


(a) Model2 (b) Model3

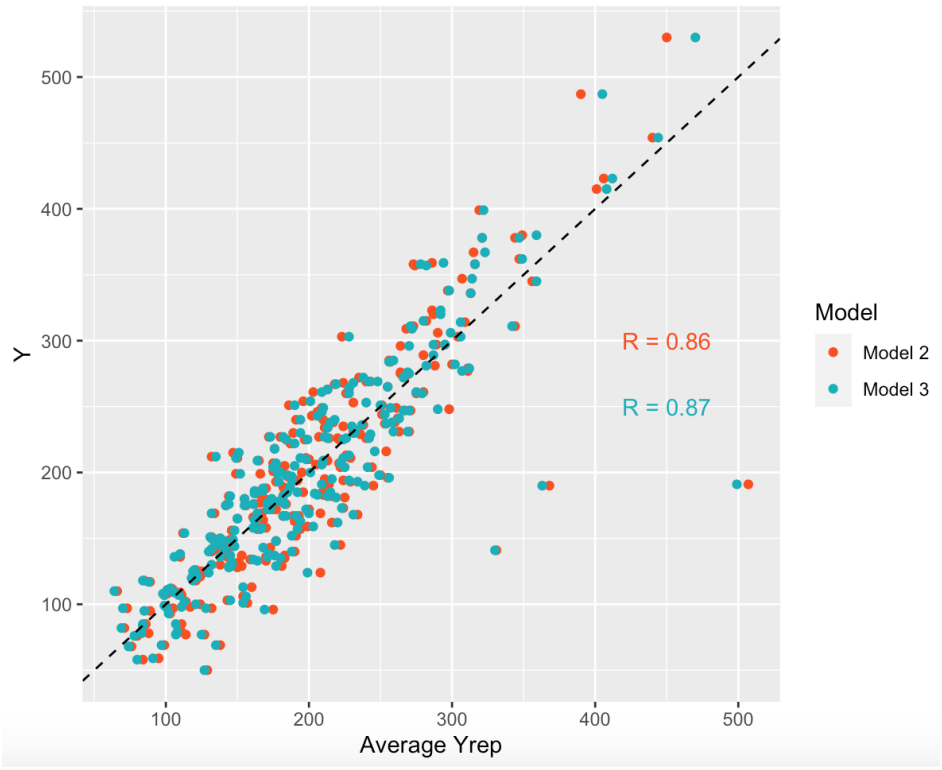
Figure 4.11: The estimated risks scatter plots of scenarios 1 - 3 of Models 2 and 3 (grid of size 500m).

Posterior predictive check

The predictive posterior checking is also carried out to verify whether the models appropriate for data. If the model fits the data well, then replicated data generated under the model should have the similar characteristic to observed data. However, the observed data at the grid level are unknown, therefore the observed data at the areal unit level are used to compare to the simulated data (grid level) from the fitted model that aggregate to the areal unit level. Figure 4.12 indicates that the observed data and the aggregated simulated data are not different since the data fairly lie on the straight line in both models and grid square sizes. There are however some outliers appearing in the plot. This is likely because the data are rescaled twice (disaggregate to the grid level and aggregate back to the areal unit level). These results suggest that the models fit the data well and the inferences are appropriate to carry out.



(a) 1,000 m



(b) 500 m

Figure 4.12: Posterior predictive model checks.

Main results

Figures 4.13 and 4.14 display the estimated disease risks from Models 2 and 3 with grid square sizes of 1,000 and 500 metres. Overall, the disease maps of grid square sizes 1,000 and 500 metres have the same pattern. Consequently, the regions with the higher risks are located on the east of Glasgow city centre and in the north, north-east and south-east such as Easterhouse, Shettieston, Possilpark, Drumchapel, Nitshill, and Castlemilk. These regions are amongst the least wealthy areas in Glasgow. In contrast, the regions of lower risks include Whitecraigs, West End, Netherlee, Newton Mearns, which are more wealthy areas. This suggests that people in poorer areas are more likely to be hospitalised for respiratory disease than those in richer areas. People in these areas are more likely to smoke, drink, have the unhealthy food consumption, and lack an exercise, which are the main causes of respiratory disease (Pampel et al., 2010; World Health Organization and others, 2007). Furthermore, I also notice that the areas with higher risks are located near motorway or a big roads. This might be because people who can afford to live anywhere, so they normally avoid living in noisy places with high pollution. The house prices in these areas are also less than areas further away from big roads.

Model 2 and Model 3 produce the similar disease maps, as can be seen in Figures 4.13 and 4.14. To quantify this similarity, Figure 4.15 presents the correlation of the estimated risks from these models. With grid square of size 1,000 metres, the correlation between the estimated risks from the models was 0.98. With grid square of size 500 metres, the correlation between the estimated risks from the models was 0.99. This means Models 2 and 3 estimate approximately similar disease risks, which is also illustrated by the maps of the estimated diffrence in disease risks between these two models in Figures 4.16 and 4.17. However, the differences between the models are bigger when the average estimated disease risks are more extreme. This can be seen from Figure 4.17 where the biggest difference occurs in poor areas. The biggest differences are for Branchton and Gilshochill for the grid square size 1,000 metres, and Castlemilk for the grid square size 500 metres.

The Mean Absolute Difference (MAD) is calculated to measure the variability between the models. It can be computed as $\frac{1}{n}\sum_{j=1}^{m}|R^3(\mathcal{G}_j) - R^2(\mathcal{G}_j)|$, where $R^i(\mathcal{G}_j), i = 2, 3.$ denotes the estimated disease risks of Model i. The MAD for grid square size 1,000 metres is 0.048, while the grid square size of 500 metres has a MAD of 0.034. The small values suggest that Model 2 and Model 3 produce very similar disease risks.

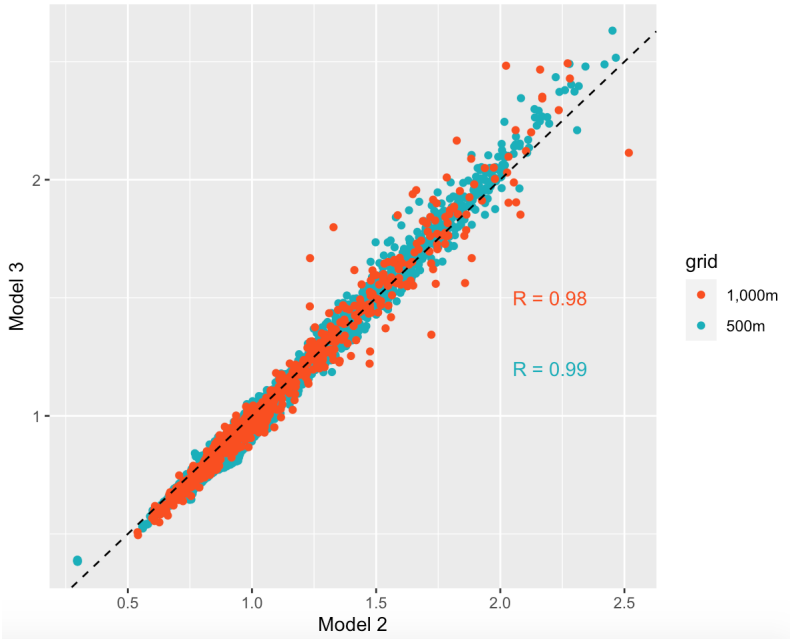


Figure 4.15: Correlation between the estimated disease risk of Models 2 and 3.

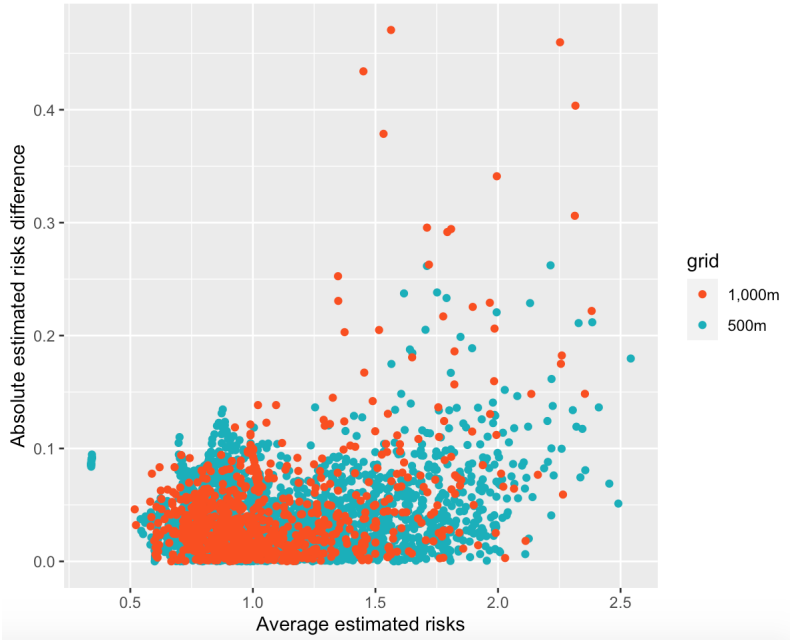
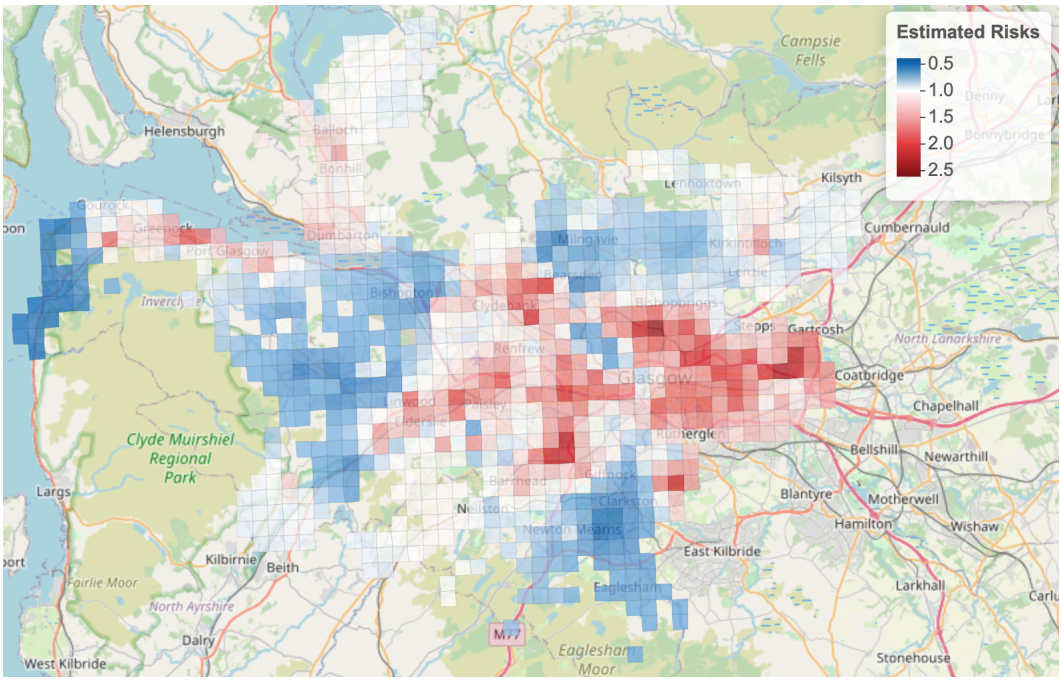
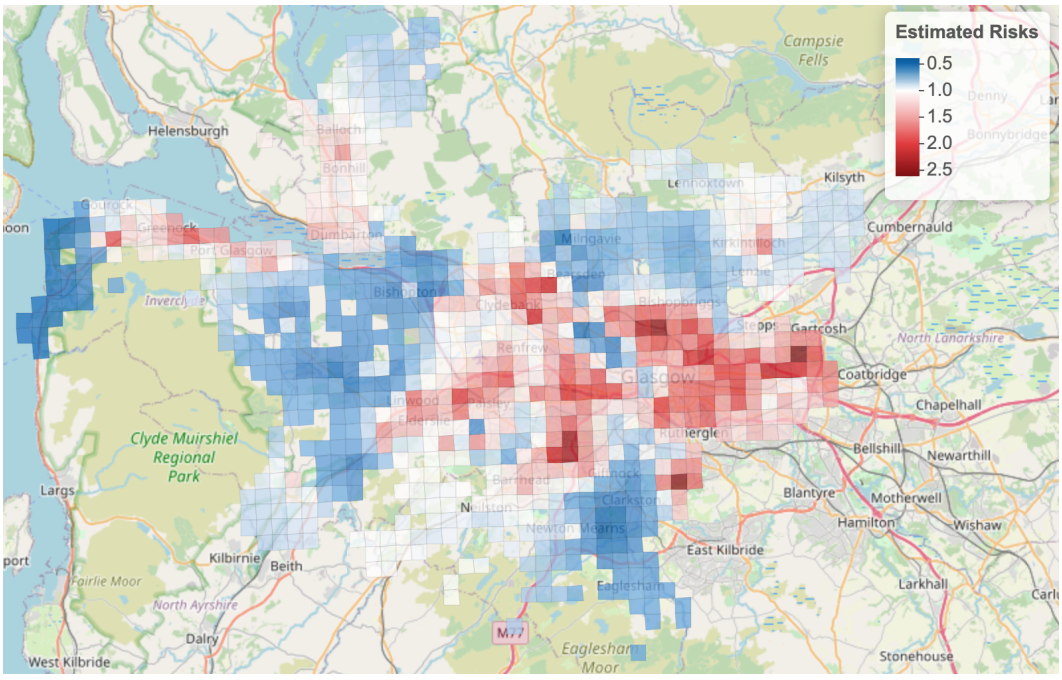


Figure 4.16: Plots of the absolute estimated disease risk difference between Models 2 and 3 versus the average of the estimated disease risk.

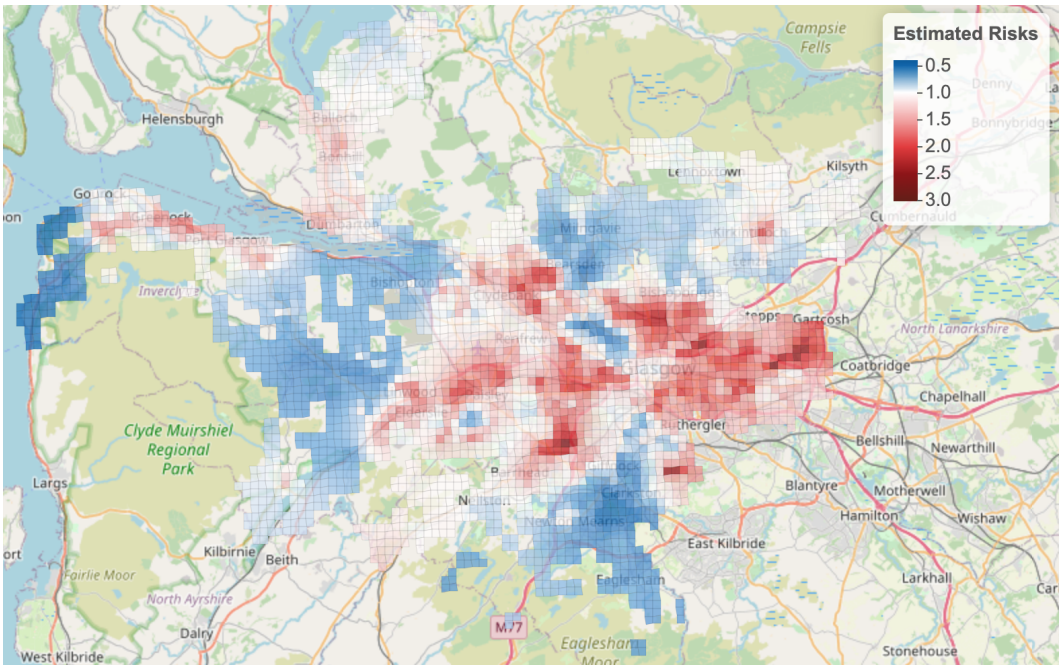


(a) Model 2 (1,000m)

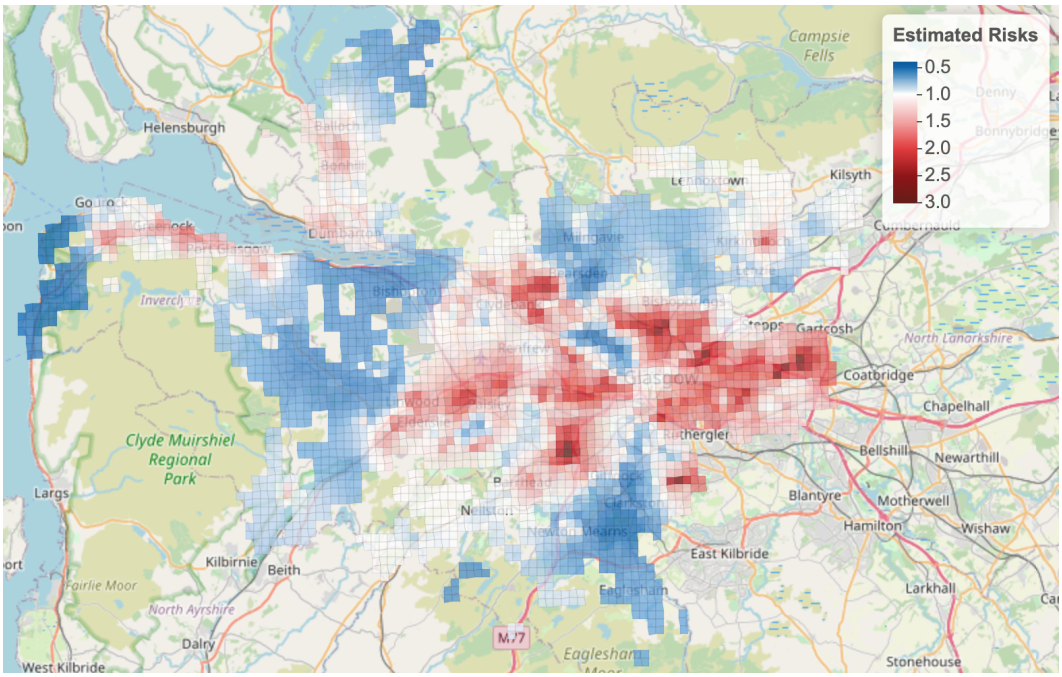


(b) Model 3 (1,000m)

Figure 4.13: Estimated disease risks from the proposed models on grid square size 1,000 metres.

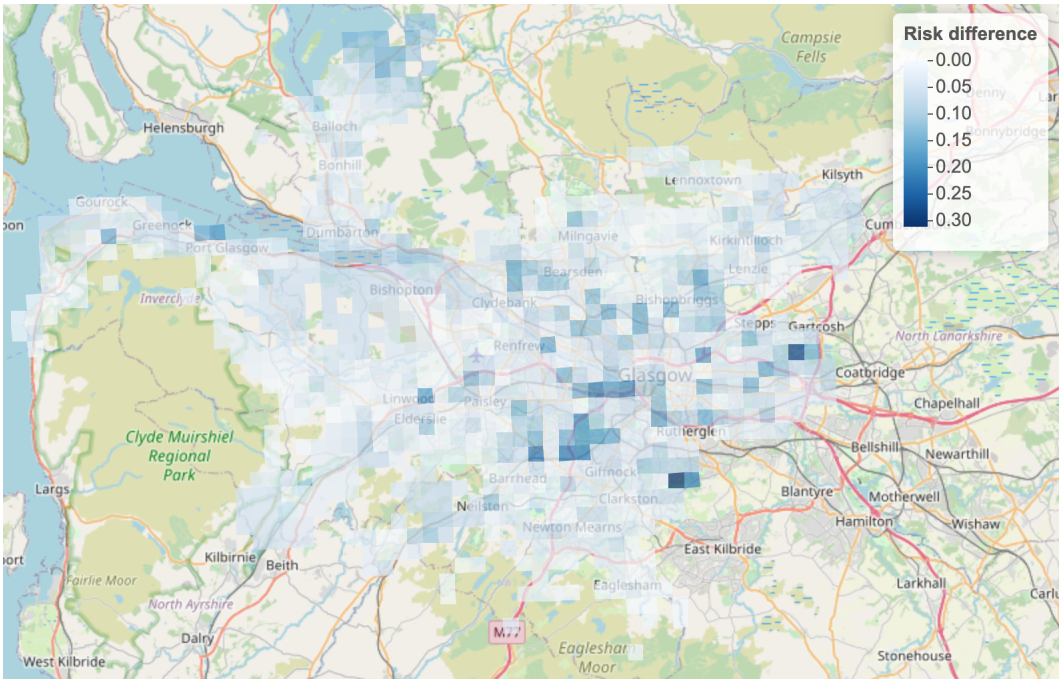


(a) Model 2 (500m)

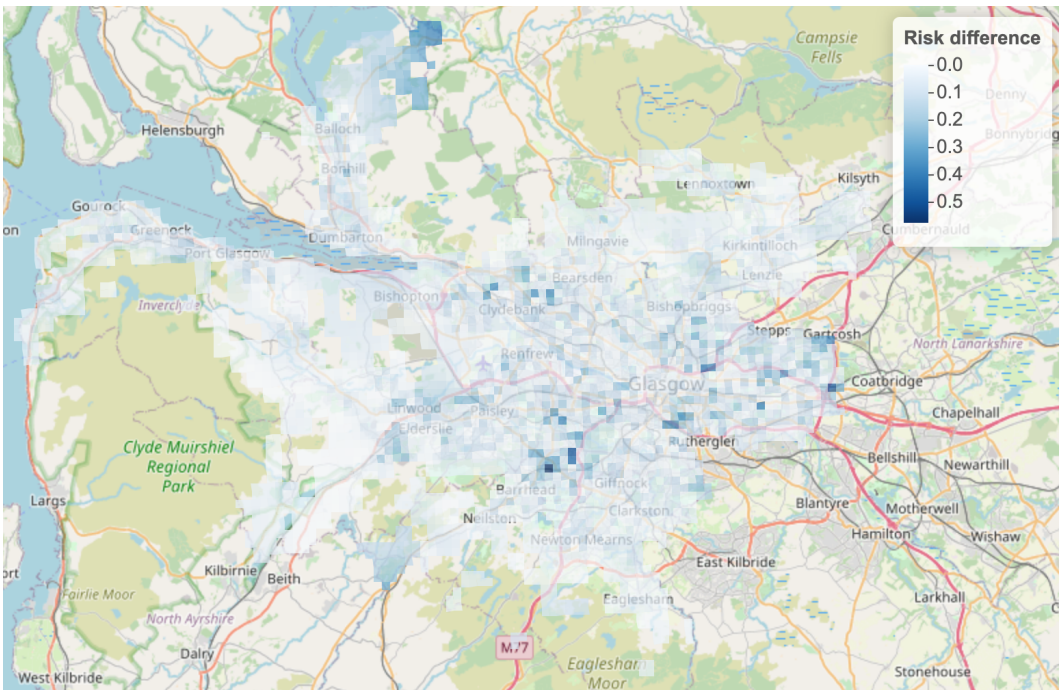


(b) Model 3 (500m)

Figure 4.14: Estimated disease risks from the proposed models on grid square size 500 metres.



(a) Grid size 1,000 metres



(b) Grid size 500 metres

Figure 4.17: The estimates disease risk difference between Models 2 and 3.

4.5 Conclusion

In this chapter, pseudo continuous inference from areal unit data has been produced. I created grid squares over the areas of Greater Glasgow and Clyde Health Board and transformed the areal unit data into grid square level data. The expected disease cases in each grid square can be computed by assuming proportionality to the population density in each grid square. The observed disease case at the grid level can be estimated via multinomial sampling, here I consider two possible approaches. The first assumes equal risk for every grid square (Model 2) and the second estimates the risks by using kriging (Model 3).

In this study the CAR model proposed by Leroux (Leroux et al., 2000) is used to estimate the spatial variation in the disease risk. The simulation study is conducted to determine how accurately the proposed model can estimate the disease risk and regression parameters at the grid level. In this simulation study, I propose three models . Model 1 is fitted to the true grid level data, Model 2 is fitted to the disaggregated data at the grid level by assuming all grid squares have constant disease risk, while Model 3 estimates the disease risks via kriging. Furthermore, the grid level data are fitted to all models with grid squares sides of lengths 1,000 and 500 metres.

The results show consistent findings across all scenarios. Model 1 performs the best for all metrics and scenarios, which is expected as it is fitted to the true grid level data. Model 2 and Model 3 produce close to unbiased estimates of regression parameters when areal level covariates are disaggregated to the grid level, while they produce biased estimates when they are true grid level covariates. This result is surprising because the true covariate data should perform better than the estimated covariate data. The reason is that disease case data at the areal unit level and the estimated covariate data are both transformed to the grid level data by the same process, therefore they have similar levels of spatial smoothness.

Model 3 has a better estimation of regression parameters corresponding to the true known covariate and the disease risks than Model 2 in terms of RMSE. This is because

Model 3 estimates the disease risk at the grid level via kriging, which leads the estimated number of disease cases being closer to the true grid level data than Model 2 which assumes disease risk at the grid level is constant. The results of the grid square 1,000 and 500 metres generally have the similar patterns, however, the grid square size of 1,000 metres is better than grid square size of 500 metres for all metrics. This is because the number of grid square size of 1,000 metres is fewer than the number of grid square size of 500 metres (853 vs 3,106 grid squares), which means that I have to estimate more grid level data than the latter, hence the results are less accurate.

The respiratory disease data in the Greater Glasgow and Clyde Health Board are applied to the proposed models (Model 2 and Model 3). Overall, the disease maps from all models fitted to data with grid square sizes of 1,000 and 500 metres have the similar pattern. The areas with higher risk are Easterhouse, Shettieston, Possilpark, Drumchapel, Nitshill, and Castlemilk which are less wealthy. In contrast, the regions of lower risk are Whitecraigs, West End, Netherlee, Newton Mearns, which are more wealthy areas. The people who live in poor areas are more likely to smoke, drink, unhealthy food consumption, and exercise less, which are the main factors responsible for the respiratory disease.

There is however, some limitations to this methodology. First, the SIRs at the grid square level were predicted by using kriging, and the SIR is an unstable estimated of risks, hence the predictions may be affected. Second, the need to store and process each of multiple imputed datasets, since the final step of this approach is combining the results from all imputed datasets. Finally, this approach allows for limited amount of uncertainty when estimating disease counts at the grid level, since they are generated for ten datasets so ten different realisations of disease counts at the grid level are used to fit the models. Therefore, in the next chapter I will introduce a new method to estimate the disease cases at the grid level which allows for higher level of uncertainty.

Chapter 5

Grid level inference with data augmentation

5.1 Introduction

In the previous chapter, a method for estimating disease risk at the grid square level was introduced. This approach consisted of two steps, the first step involved estimating disease cases at the grid level via multiple imputation, and the second step involved fitting the conditional autoregressive model proposed by Leroux et al. (2000) to these imputed data. I considered two approaches for imputing the disease cases; the first assumed equal risk across all the grid squares and the second estimated the disease risks via kriging. The latter approach performed better and therefore in this chapter only the second approach will be used as a comparator to that proposed here.

There are, however, some disadvantages to this method of estimating disease cases at the grid level. I estimated the disease cases via a single multiple imputation step prior to fitting a spatial model and repeated these steps for ten times then combined the results in order to estimate disease risk at the grid level. Therefore, there was limited level of uncertainty in the disease counts when estimating the model parameters. Furthermore, disease risk at the grid level needs to be estimated in a preliminary manner before I estimate the numbers of disease cases in each grid square via multiple imputation. Thus if these initial grid level estimated risks are not accurate then the results of

model fitting are unlikely to be accurate either. This could easily be the case in some scenarios, because the SIR that is then kriged is an unstable estimate of disease risk. In this chapter, I will attempt to address these issues by using a data augmentation approach.

Data augmentation was proposed by Tanner and Wong (1987), and is generally used to handle censored or unobserved data. It is commonly applied in Bayesian statistics, especially in the application of Markov chain Monte Carlo (MCMC) simulation (Neal and Kypraios, 2015). The general idea of data augmentation in this context is based on two iterative steps. In the first step, given the current values of the parameters of interest and the disease counts at the areal unit level, I estimate the disease counts at the grid square level by drawing from a multinomial distribution. In the second step, I update the parameters from their full conditional posterior distribution based on the newly estimated grid level disease counts from the previous step. These two steps are repeated within an MCMC algorithm to allow for uncertainty in both the data and the parameters. More detail about this data augmentation algorithm will be presented in Section 5.2.1.

The remainder of this chapter will be organised as follows. Section 5.2 outlines the proposed model, and Section 5.3 uses simulated data to test this proposed model against the model from the previous chapter. Section 5.4 presents an application of this methodology, based on respiratory hospital admissions in the Greater Glasgow and Clyde Health Board from January 2015 to December 2016. Finally, Section 5.5 discusses the advantages and disadvantages of this methodology.

5.2 Methodology

5.2.1 Data augmentation

Recall that grid level inference is based on the model

$$\begin{aligned} Y(\mathcal{G}_j) &\sim \text{Poisson}[e(\mathcal{G}_j)R(\mathcal{G}_j)] \quad \text{for} \quad j = 1, ..., m \\ \ln(R(\mathcal{G}_j)) &= \mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j) \\ \phi(\mathcal{G}_j)|\boldsymbol{\phi}(\mathcal{G}_{-j}) &\sim \text{N}\left(\frac{\rho \sum_{k=1}^m w_{kj}\phi(\mathcal{G}_k)}{\rho \sum_{k=1}^m w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{k=1}^m w_{kj} + 1 - \rho}\right) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho &\sim \text{Uniform}(0, 1), \end{aligned} \tag{5.2.1}$$

where $e(\mathcal{G}_j)$ and $R(\mathcal{G}_j)$ denote expected disease cases and disease risk at grid square $\mathcal{G}_j$ respectively, $\mathbf{x}(\mathcal{G}_j)$ is the vector of covariates for grid square $\mathcal{G}_j$, $\boldsymbol{\beta}$ is the set of grid level regression parameters and $\phi(\mathcal{G}_j)$ is the spatially autocorrelated random effect for grid square $\mathcal{G}_j$. $\boldsymbol{Y}(\mathcal{G}) = [Y(\mathcal{G}_1), ..., Y(\mathcal{G}_m)]$ is the vector of disease counts at the grid square level, which can be estimated by multinomial sampling steps. Let us denote $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ as the number of disease counts in the intersection area between $\mathcal{A}_i$ and $\mathcal{G}_j$. Then $Y(\mathcal{G}_j)$ can be computed by

$$Y(\mathcal{G}_j) = \sum_{i=1}^{n} Y(\mathcal{A}_i \cap \mathcal{G}_j). \tag{5.2.2}$$

We can estimate $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ by partitioning the disease counts in area $\mathcal{A}_i$ across the m grid squares intersections $\{\mathcal{A}_i \cap \mathcal{G}_1, ..., \mathcal{A}_i \cap \mathcal{G}_m\}$ using a multinomial sampling. Therefore given $Y(\mathcal{A}_i)$, $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ can be estimated as follows

$$[Y(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y(\mathcal{A}_i \cap \mathcal{G}_m)] \sim \text{Multinomial}(n = Y(\mathcal{A}_i)|\omega_{i1}, ..., \omega_{im}). \tag{5.2.3}$$

In the final step, combine $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ via (5.2.2) to estimate $Y(\mathcal{G}_j)$ for each grid square $\mathcal{G}_j$. However $Y(\mathcal{A}_i \cap \mathcal{G}_j)$ should be drawn multiple times to reduce the variability in the sampled data. Therefore the data are drawn for L iterations from the multinomial step 5.2.3, $[Y^{(l)}(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y^{(l)}(\mathcal{A}_i \cap \mathcal{G}_m)]$ for $l = 1, ..., L$ and then take the mean;

$$Y(\mathcal{G}_j) = \frac{1}{L}\sum_{l=1}^{L} Y^{(l)}(\mathcal{G}_j) \qquad \text{for} \quad j = 1, ..., m. \tag{5.2.4}$$

There are some issues occur in the estimation of $Y(\mathcal{G}_j)$, which are $Y(\mathcal{G}_j)$ is not necessarily an integer and the sum of disease count at the grid level is not necessarily equal to the areal unit level, i.e. $\sum_{j=1}^{m} Y(\mathcal{G}_j) \neq \sum_{i=1}^{n} Y(\mathcal{A}_i)$. The methodology to overcome these problems are outlined in Section 4.2.2. However, the multinomial samples cannot be drawn if we do not know the probability (ω_{ij}) of each disease case in area $\mathcal{A}_i$ that lie in the intersection area $\mathcal{A}_i \cap \mathcal{G}_j$. The probability can be defined similar to Chapter 4 as follows;

$$\omega_{ij} = \frac{e(\mathcal{G}_j)\hat{R}(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} e(\mathcal{G}_k)\hat{R}(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}}, \tag{5.2.5}$$

where $\hat{R}(\mathcal{G}_j)$ is the estimated risk at the grid level which can be estimated via kriging. Full details of kriging method are outlined in Sections 2.3.5 and 4.2.3.

In the previous chapter, I estimated disease counts at the grid square level, $\boldsymbol{Y}(\mathcal{G}) = [Y(\mathcal{G}_1), ..., Y(\mathcal{G}_m)]$ via multiple imputation steps, and then based on these data, fitted a spatial model to estimate disease risks at the grid level, $\boldsymbol{R}(\mathcal{G}) = [R(\mathcal{G}_1), ..., R(\mathcal{G}_m)]$. However, this approach allows for a limited amount of uncertainty when estimating the model parameters based on the imputed values of $\boldsymbol{Y}(\mathcal{G})$, since I estimate $Y(\mathcal{G}_j)$ by multiple imputation and kriging once and then keep these estimates for every MCMC iteration when fitting (5.2.1) for each dataset, and then pool the results for model inference. Additionally, the model inference is dependent on the accuracy of the original estimates of $\hat{\boldsymbol{R}}(\mathcal{G})$ obtained by kriging. If the model is fitted to less accurate data, then it follows that the model inference will be less reliable.

One way to overcome these problems is via data augmentation. This approach allows the model to update the disease count in each grid square, $Y(\mathcal{G}_j)$, within the MCMC algorithm, which means that $\boldsymbol{Y}(\mathcal{G})$ will be updated when estimating the model parameters at the same time. This allows uncertainty when estimating $Y(\mathcal{G}_j)$, and should improve the accuracy of the estimates of $\boldsymbol{R}(\mathcal{G})$ and $\boldsymbol{Y}(\mathcal{G})$. In order to obtain more sta-

ble estimates of $\boldsymbol{R}(\mathcal{G})$, I only update $\boldsymbol{Y}(\mathcal{G})$ at every K MCMC iterations, and average over these when carrying out the multinomial step. This has the additional benefit of reducing computational time, because it reduces the number of multinomial draws needed. The data augmentation algorithm is as follows

Data augmentation algorithm

1. Generate initial values for all model parameters, $\boldsymbol{\Theta}^{(0)} = [\boldsymbol{\beta}^{(0)}, \boldsymbol{\phi}^{(0)}(\mathcal{G}), \tau^{2(0)}, \rho^{(0)}]$ and disease counts at the grid level, $\boldsymbol{Y}^{(0)}(\mathcal{G})$.

2. Iterate the following steps for $t = 1, ..., T$ iterations.

   a) Update each model parameter in turn via MCMC steps using Gibbs sampling or Metropolis-Hastings steps.

   - Update $\boldsymbol{\beta}^{(t)}|\boldsymbol{Y}^{(t-1)}(\mathcal{G}), \boldsymbol{\phi}^{(t-1)}(\mathcal{G})$.
   - Update $\boldsymbol{\phi}^{(t)}(\mathcal{G})|\boldsymbol{Y}^{(t-1)}(\mathcal{G}), \boldsymbol{\beta}^{(t)}, \tau^{2(t-1)}, \rho^{(t-1)}$.
   - Update $\tau^{2(t)}|\boldsymbol{\phi}^{(t)}(\mathcal{G}), \rho^{(t-1)}$.
   - Update $\rho^{(t)}|\boldsymbol{\phi}^{(t)}(\mathcal{G}), \tau^{2(t)}$.

   b) If t is a multiple of K, update disease counts at the grid level $\boldsymbol{Y}^{(t)}(\mathcal{G})|\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}(\mathcal{G}), \tau^{2(t)}, \rho^{(t)}$ via multinomial sampling as

$$Y^{(t)}(\mathcal{G}_j) = \sum_{i=1}^{n} Y^{(t)}(\mathcal{A}_i \cap \mathcal{G}_j), \tag{5.2.6}$$

   where

$$\left[Y^{(t)}(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y^{(t)}(\mathcal{A}_i \cap \mathcal{G}_m)\right] \sim \text{Multinomial}(n = Y(\mathcal{A}_i)|\omega_{i1}, ..., \omega_{im}), \tag{5.2.7}$$

   the weight ω_{ij} is the probability of each disease case in region $\mathcal{A}_i$ occurring in the intersection area between grid square $\mathcal{G}_j$ and region $\mathcal{A}_i$, $a(\mathcal{A}_i \cap \mathcal{G}_j)$. Here, I assume that ω_{ij} is dependent on the expectation of the disease count in the intersection area ($\omega_{ij} \propto \mathbb{E}[Y(\mathcal{A}_i \cap \mathcal{G}_j)]$). However, $\mathbb{E}[Y(\mathcal{A}_i \cap \mathcal{G}_j)]$ is unknown, therefore I estimate it by

$$\mathbb{E}[Y(\mathcal{A}_i \cap \mathcal{G}_j)] \approx \mathbb{E}[Y(\mathcal{G}_j)] \times \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)},$$

which is proportional to the area of grid square $\mathcal{G}_j$ which lies in region $\mathcal{A}_i$. From the model 5.2.1, I know that $\mathbb{E}[Y(\mathcal{G}_j)] = e(\mathcal{G}_j)R(\mathcal{G}_j)$, so that

$$\mathbb{E}[Y(\mathcal{A}_i \cap \mathcal{G}_j)] \approx e(\mathcal{G}_j)R(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}.$$

As before, $e(\mathcal{G}_j)$ and $R(\mathcal{G}_j)$ are the expected disease count and the disease risk in grid square $\mathcal{G}_j$ respectively. Therefore I have

$$\omega_{ij} = \frac{e(\mathcal{G}_j)\tilde{R}(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} e(\mathcal{G}_k)\tilde{R}(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}}, \tag{5.2.8}$$

the denominator is used to ensure that $\sum_{j=1}^{m} \omega_{ij} = 1$ for all $i = 1, ..., n$. $\tilde{R}(\mathcal{G}_j)$ is obtained by averaging over the previous K MCMC iterations. Here let $R^{(t)}(\mathcal{G}_j)$ denote the current estimate of disease risk in grid square $\mathcal{G}_j$ which can be estimated as follows

$$\boldsymbol{R}^{(t)}(\mathcal{G}) = [R^{(t)}(\mathcal{G}_1), ..., R^{(t)}(\mathcal{G}_m)] = \exp[\boldsymbol{X}(\mathcal{G})^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{\phi}^{(t)}(\mathcal{G})]. \tag{5.2.9}$$

Then I can compute the average of the last K $R^{(t)}(\mathcal{G})$ values, which is given by

$$\tilde{R}(\mathcal{G}_j) = \frac{1}{K}\sum_{r=t-K+1}^{t} R^{(r)}(\mathcal{G}_j). \tag{5.2.10}$$

Unlike the previous imputation approach, here I simultaneously estimate both the data $Y(\mathcal{G}_j)$ and the parameters controlling the spatial surface $\phi(\mathcal{G}_j)$, which allows for uncertainty in the former. However, initial results showed that to guarantee a spatially smooth surface I needed to fix $\rho = 1$ to enforce strong spatial smoothness for the disease risk. Note that the Leroux CAR model with $\rho = 1$ corresponds to the Intrinsic CAR model proposed by Besag et al. (1991).

I carried out a preliminary simulation study and established that for a relatively small number of simulated datasets, τ^2 is massively overestimated, leading to very poor disease risk estimation at the grid level, because the estimated disease risks have too

much variation. This is due to a lack of identifiability, since this approach requires me to update $Y(\mathcal{G}_j)$ and model parameter $\phi(\mathcal{G}_j)$ at the same time and some parameters converge to the wrong values. This is different from the multiple imputation approach I proposed in the previous chapter, where $Y(\mathcal{G}_j)$ is updated independently prior to estimating the model parameters.

To avoid this identifiability problem, I fix τ^2 in (5.2.1), thus ensuring that the values of $\phi(\mathcal{G}_j), \tilde{R}(\mathcal{G})$, and the data augmented disease count, $Y(\mathcal{G}_j)$ remain stable. In order to find an appropriate value for τ^2, I use an empirical Bayes approach (Casella, 1985). Consider the original grid level model (5.2.1)

$$\begin{aligned} Y(\mathcal{G}_j) &\sim \text{Poisson}[e(\mathcal{G}_j)R(\mathcal{G}_j)] \quad \text{for} \quad j = 1, ..., m \\ \ln[R(\mathcal{G}_j)] &= \mathbf{x}(\mathcal{G}_j)^\top \boldsymbol{\beta} + \phi(\mathcal{G}_j) \\ \boldsymbol{\phi}(\mathcal{G}) &\sim \text{N}(\mathbf{0}, \tau^2 \boldsymbol{Q}^-), \end{aligned} \tag{5.2.11}$$

where $\boldsymbol{\phi}(\mathcal{G}) \sim \text{N}(\mathbf{0}, \tau^2\boldsymbol{Q}^-)$ is the multivariate analogue of the univariate conditional distributions given by (5.2.1) with $\rho = 1$. Here $\boldsymbol{Q} = \text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}$, and $\boldsymbol{Q}^-$ is the generalised inverse as $\boldsymbol{Q}$ is singular. Thus I replace the singular matrix $\boldsymbol{Q}$ with the invertible matrix, $\tilde{\boldsymbol{Q}}$ (Lee et al., 2014) by adding a small constant ($\epsilon = 0.001$) onto the diagonal terms of $\boldsymbol{Q}$ to make it invertible, i.e. $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} + \epsilon\mathbf{I}$. This model has expectation,

$$\begin{aligned} \mathbb{E}[\boldsymbol{Y}(\mathcal{G})] &= \boldsymbol{e}(\mathcal{G})\boldsymbol{R}(\mathcal{G}) \\ &= \boldsymbol{e}(\mathcal{G})\exp[\boldsymbol{X}(\mathcal{G})\boldsymbol{\beta} + \boldsymbol{\phi}(\mathcal{G})]. \end{aligned}$$

Thus

$$\ln\left(\mathbb{E}\left[\frac{\boldsymbol{Y}(\mathcal{G})}{\boldsymbol{e}(\mathcal{G})}\right]\right) = \boldsymbol{X}(\mathcal{G})\boldsymbol{\beta} + \boldsymbol{\phi}(\mathcal{G}).$$

I therefore see that,

$$\ln\left(\mathbb{E}\left[\frac{\boldsymbol{Y}(\mathcal{G})}{\boldsymbol{e}(\mathcal{G})}\right]\right) \sim \text{approx} \quad \text{N}(\boldsymbol{X}(\mathcal{G})\boldsymbol{\beta}, \tau^2\tilde{\boldsymbol{Q}}^{-1}). \tag{5.2.12}$$

The values of $\ln\left(\mathbb{E}\left[\frac{\boldsymbol{Y}(\mathcal{G})}{\boldsymbol{e}(\mathcal{G})}\right]\right)$ are the logged expected SIR values for the m grid squares, which from the previous chapter (multiple imputation approach) can be estimated via kriging. Thus I can denote the vector of logged expected SIR for all grid square

locations based on kriging by $\boldsymbol{Z}(\mathcal{G}) = \ln[\mathbf{SIR}(\mathcal{G})]$. The maximum likelihood estimates of $\boldsymbol{\beta}$ and τ^2 from 5.2.12 are therefore given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [\boldsymbol{X}^\top(\mathcal{G})\tilde{\boldsymbol{Q}}\boldsymbol{X}(\mathcal{G})]^{-1}\boldsymbol{X}^\top(\mathcal{G})\tilde{\boldsymbol{Q}}\boldsymbol{Z}(\mathcal{G}) \\ \hat{\tau}^2 &= \frac{1}{m-p}[\boldsymbol{Z}(\mathcal{G}) - \boldsymbol{X}(\mathcal{G})\hat{\boldsymbol{\beta}}]^\top\tilde{\boldsymbol{Q}}[\boldsymbol{Z}(\mathcal{G}) - \boldsymbol{X}(\mathcal{G})\hat{\boldsymbol{\beta}}].\end{aligned} \tag{5.2.13}$$

Thus, when fitting (5.2.1), I fix $\tau^2 = \hat{\tau}^2$ to prevent unrealistic values of $R(\mathcal{G}_j)$ being estimated.

5.3 Simulation study

5.3.1 Aim

A simulation study is conducted to establish the efficacy of the data augmentation modelling approach outlined in the previous section, and to compare the model to the multiple imputation modelling approach outlined in the previous chapter.

5.3.2 General approach

This simulation study follows a similar approach to the previous chapter, and comprises four main steps. First, I generate disease counts, expected disease cases, and covariates $[Y(\mathcal{G}_j), e(\mathcal{G}_j), \mathbf{x}(\mathcal{G}_j)]$ at the grid level. Second, the generated grid level data in step one are aggregated to the areal unit level to reflect the type of data that normally occur in practice. Then, the models are fitted under eight different scenarios to reflect the variety of real life scenarios that may arise. Finally, I repeat steps one to three for 100 simulated datasets and summarise the results to measure the accuracy of the models.

5.3.3 Grid level data generation

Grid level data are generated for the Greater Glasgow and Clyde Health Board region with grid squares of sizes 1,000 and 500 metres, giving 853 and 3,106 grid squares respectively (after removing grid squares where no people live). The data generation is carried out in the same manner as described in Section 4.3.3. Specifically, the data are generated from model (5.2.1), which involves a disease count, $Y(\mathcal{G}_j)$, an ex-

Table 5.1: The scenarios used in the simulation study.

Scenario	ρ	τ^2	ψ
1	0.99	0.01	0.01
2	0.99	0.01	0.05
3	0.5	0.01	0.01
4	0.5	0.01	0.05
5	0.99	0.05	0.01
6	0.99	0.05	0.05
7	0.5	0.05	0.01
8	0.5	0.05	0.05

pected disease count, $e(\mathcal{G}_j)$, disease risk in grid square $\mathcal{G}_j$, $R(\mathcal{G}_j)$, regression parameters (β_1, β_2), covariates $[\mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j)]$, spatial random effect, $\phi(\mathcal{G}_j)$ and additional parameters (ρ, τ^2).

The expected disease count for grid square $\mathcal{G}_j$ can be generated via $e(\mathcal{G}_j) = \psi P(\mathcal{G}_j)$, where ψ is the proportion of people who have the disease event in grid square $\mathcal{G}_j$ and $P(\mathcal{G}_j)$ is the adjusted population for grid square $\mathcal{G}_j$. Note that the value of ψ is varied in the simulation study to see how disease prevalence affects model performance. Next, I want to generate $R(\mathcal{G}_j)$, which means I have to first generate $\beta_1, \beta_2, \mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j), \phi(\mathcal{G}_j)$ and (ρ, τ^2). Here I fix β_1 and β_2 at 0.1 as in the previous chapter, and $\mathrm{x}_1(\mathcal{G}_j)$ and $\mathrm{x}_2(\mathcal{G}_j)$ are generated from normal distributions with mean zero and variance one, $[\mathrm{x}_1(\mathcal{G}_j) \sim \mathrm{N}(0,1), \mathrm{x}_2(\mathcal{G}_j) \sim \mathrm{N}(0,1)]$. Finally, the vector of spatial random effects, $\boldsymbol{\phi}(\mathcal{G})$ is generated from a multivariate normal distribution with mean zero and variance $\tau^2 \boldsymbol{Q}^{-1}$, where $\boldsymbol{Q} = \rho[\mathrm{diag}(\boldsymbol{W}\boldsymbol{1} - \boldsymbol{W}] + (1-\rho)\boldsymbol{I}$, and $\boldsymbol{W}$ is a neighbourhood matrix at the grid level. This model corresponds to the random effects from the conditional autoregressive (CAR) model proposed by Leroux et al. (2000), which is the same model I used in the previous chapter. In this simulation study τ^2 and ρ are varied as well as ψ. The 100 datasets are generated under the same set of scenarios as the previous chapter, these are outlined in Table 5.1

5.3.4 Data aggregation

Next, I have to aggregate the grid level data, $[Y(\mathcal{G}_j), e(\mathcal{G}_j), \mathrm{x}_1(\mathcal{G}_j), \mathrm{x}_2(\mathcal{G}_j)]$ to the areal unit level $[Y(\mathcal{A}_i), e(\mathcal{A}_i), \mathrm{x}_1(\mathcal{A}_i)]$, to reflect the data I typically have in real applications.

Note that I do not aggregate $\mathrm{x}_2(\mathcal{G}_j)$ to the areal level, since it represents real data which are available at the grid level e.g. air pollution concentrations. More detail of the aggregation methods for these grid level data is provided in Section 4.3.4.

5.3.5 Fitting the model

I compare three different models in this chapter. In Model 1, I fit a CAR model to the true grid level data $[Y(\mathcal{G})_j, e(\mathcal{G}_j), \mathrm{x}_1(\mathcal{G}_i), \mathrm{x}_2(\mathcal{G}_j)]$. This acts as a reference model to compare to the others. Next, Model 3 from the previous chapter is fitted to the estimated disaggregated grid level data $[\tilde{Y}(\mathcal{G}_j), \tilde{e}(\mathcal{G}_j), \tilde{\mathrm{x}}_1(\mathcal{G}_j)]$ and the grid level covariate $[\mathrm{x}_2(\mathcal{G}_j)]$. It was the best performing multiple imputation model, and thus acts as a comparison between multiple imputation and data augmentation. Finally, I fit the extended model with the data augmentation approach described in this chapter. I consider two variants of the data augmentation approach; Model 4 is the standard approach, where τ^2, the variance of the spatial random effects, is estimated when fitting the model, while Model 5 fixes τ^2 using empirical Bayes. In summary, the four models are outlined as follows (note that the numbering used maintains consistency with the previous chapter)

- Model 1 - fit the model to the true grid level data (reference model).
- Model 3 - fit the model to the disaggregated data at the grid level with disease risk, $\hat{R}(\mathcal{G}_j)$ estimated via kriging.
- Model 4 - fit the data augmentation model without fixing τ^2.
- Model 5 - fit the data augmentation model with τ^2 fixed.

Model 1 is included as a reference model to compare the performance to other models. It is fitted to the true grid level data, and therefore should perform the best of the four models. I am primarily interested in which of Models 3, 4, or 5 perform best, and how close their estimates are to those from Model 1.

In this study, I generate $r = 100$ datasets to estimate the regression parameters and disease risk at the grid level. Parameters β_1, β_2, τ^2 and $\boldsymbol{\phi}(\mathcal{G})$ are estimated by the

posterior median obtained via the MCMC samples for each simulated data set. In addition, the uncertainty associated with these estimates is represented by the upper and lower limits of the 95% credible interval for each parameter, which are respectively the 2.5th and 97.5th percentile from the MCMC samples for each dataset. Inference for each dataset is based on 200,000 MCMC samples where I discard the first 50,000 samples as burn-in. I thin the remaining 150,000 samples by a factor of 15, leaving a total of 10,000 samples for model inference. Convergence diagnostic is done by using the method introduced by Geweke et al. (1991) and traceplots.

5.3.6 Summarising the results

To measure the performance of the four models in terms of estimating regression parameters and disease risk at the grid level, I use the four metrics outlined in Section 3.3.6 which are bias, root mean square error (RMSE), credible interval (CI) coverage, and average width of CI. More detail about these four metrics is provided in Section 4.3.6.

5.3.7 Simulation results

There are 100 simulated datasets in this study, under two different grid square sizes; with sides of lengths 1,000 and 500 metres. The simulation results for the grid square with 1,000 metres length are shown in Tables 5.2, 5.3, and 5.4, while those for the grid square with 500 metres length are shown in Table 5.5, 5.6, and 5.7. Note that I have run extra datasets for Model 4 to obtain 100 converged datasets since the results for Model 4 are unstable and it produces some extreme values for the estimated risks at the grid level due to some MCMC chains do not converge.

Table 5.2: Results from the simulation study for the regression parameter $\beta_1 = 0.1$ with the estimated covariate at the grid level $x_1(\mathcal{G}_j)$.

Metric	Scenario	Model 1	Model 3	Model 4	Model 5
Bias	1	-0.0006	0.0074	0.0127	0.0124
	2	-0.0005	0.0119	0.0172	0.0094
	3	0.0007	0.0048	0.0081	0.0039
	4	0.0004	0.0067	0.0127	0.0093
	5	0.0002	0.0178	0.0150	0.0087
	6	0.0004	0.0228	0.0261	0.0105
	7	0.0013	0.0088	0.0148	0.0075
	8	0.0020	0.0095	0.0270	0.0148
RMSE	1	0.0091	0.0241	0.0218	0.0288
	2	0.0050	0.0162	0.0246	0.0209
	3	0.0118	0.0223	0.0258	0.0277
	4	0.0059	0.0169	0.0220	0.0214
	5	0.0111	0.0277	0.0324	0.0316
	6	0.0075	0.0318	0.0348	0.0313
	7	0.0110	0.0247	0.0331	0.0335
	8	0.0094	0.0229	0.0440	0.0309
CI coverage	1	0.899	0.98	0.84	0.81
	2	0.98	0.86	0.68	0.64
	3	0.90	0.98	0.90	0.86
	4	0.94	0.86	0.69	0.58
	5	0.96	1.00	0.87	0.81
	6	0.96	0.72	0.70	0.57
	7	1.00	0.96	0.84	0.78
	8	0.964	0.82	0.63	0.47
Average width	1	0.0403	0.0977	0.0862	0.0784
	2	0.0211	0.0553	0.0473	0.0317
	3	0.0399	0.0942	0.0816	0.0761
	4	0.0220	0.0515	0.0453	0.0358
	5	0.0469	0.1188	0.1015	0.0810
	6	0.0318	0.0761	0.0792	0.0452
	7	0.0493	0.1096	0.0949	0.0813
	8	0.0369	0.0698	0.0796	0.0398

Table 5.3: Results from the simulation study for the regression parameter $\beta_2 = 0.1$ with the true covariate at the grid level $x_2(\mathcal{G}_j)$.

Metric	Scenario	Model 1	Model 3	Model 4	Model 5
Bias	1	-0.0017	-0.0657	0.0011	-0.0030
	2	-0.0012	-0.0685	0.0001	-0.0020
	3	0.0001	-0.0628	0.0004	0.0009
	4	0.0003	-0.0687	0.0008	-0.0002
	5	0.0024	-0.0676	0.0014	0.0029
	6	-0.0004	-0.0719	0.0032	-0.0010
	7	-0.0020	-0.0659	0.0010	-0.0026
	8	0.0002	-0.0729	0.0016	-0.0002
RMSE	1	0.0101	0.0661	0.0154	0.0176
	2	0.0060	0.0686	0.0089	0.0108
	3	0.0103	0.0631	0.0159	0.0206
	4	0.0054	0.0688	0.0107	0.0105
	5	0.0135	0.0679	0.0181	0.0208
	6	0.0088	0.0720	0.0173	0.0156
	7	0.0127	0.0663	0.0199	0.0210
	8	0.0090	0.0731	0.0149	0.0175
CI coverage	1	0.95	0	0.94	0.87
	2	0.93	0	0.91	0.90
	3	0.97	0	0.93	0.87
	4	0.97	0	0.84	0.75
	5	0.91	0	0.92	0.80
	6	0.91	0	0.84	0.74
	7	0.92	0	0.87	0.81
	8	0.95	0	0.87	0.54
Average width	1	0.0402	0.0496	0.0579	0.0556
	2	0.0212	0.0249	0.0303	0.0262
	3	0.0398	0.0495	0.0582	0.0561
	4	0.0220	0.0248	0.0294	0.0255
	5	0.0471	0.0543	0.0647	0.0574
	6	0.0316	0.0295	0.0428	0.0292
	7	0.0489	0.0528	0.0618	0.0570
	8	0.0373	0.0289	0.0426	0.0275

Table 5.4: Results from the simulation study for the disease risk at the grid level $R(\mathcal{G}_j)$.

Metric	Scenario	Model 1	Model 3	Model 4	Model 5
Bias	1	-0.0005	-0.0069	-0.0035	-0.0025
	2	0.0006	-0.0019	-0.0020	-0.0008
	3	0.0004	-0.0012	-0.0021	-0.0006
	4	0.0003	-0.0019	-0.0027	0.0002
	5	0.0010	-0.0123	-0.0106	-0.0025
	6	0.0015	-0.0081	-0.0054	-0.0022
	7	-0.0011	-0.0054	-0.0058	-0.0040
	8	0.0006	-0.0023	-0.0063	-0.0009
RMSE	1	0.0819	0.1423	0.1225	0.1825
	2	0.0658	0.1376	0.1158	0.1292
	3	0.0719	0.1345	0.1206	0.2037
	4	0.0641	0.1362	0.1164	0.1352
	5	0.1507	0.1976	0.1861	0.2067
	6	0.1216	0.1899	0.1713	0.1770
	7	0.1450	0.1926	0.1840	0.2145
	8	0.1261	0.1900	0.1748	0.1843
CI coverage	1	0.94	0.72	0.75	0.41
	2	0.95	0.62	0.69	0.37
	3	0.93	0.63	0.69	0.35
	4	0.95	0.55	0.64	0.29
	5	0.95	0.82	0.82	0.50
	6	0.95	0.71	0.83	0.47
	7	0.94	0.71	0.70	0.40
	8	0.95	0.65	0.80	0.34
Average width	1	0.3052	0.2995	0.2791	0.1595
	2	0.2468	0.2401	0.2308	0.1151
	3	0.2640	0.2403	0.2450	0.1383
	4	0.2439	0.2031	0.2130	0.0928
	5	0.5596	0.5134	0.4746	0.2625
	6	0.4444	0.3840	0.4433	0.2127
	7	0.5443	0.4050	0.3835	0.2089
	8	0.4741	0.3468	0.4348	0.1558

The results for all metrics and models for the grid square with size 1,000 metres are presented in Table 5.2, 5.3, and 5.4. Overall Model 1 performs the best for regression parameters $[\beta_1, \beta_2]$, and disease risk $[R(\mathcal{G}_j)]$, which is not surprising since it is fitted to the true grid level data and is thus expected to be the best model. Model 1 produces unbiased estimates across all scenarios, and it also produces the smallest RMSE, which suggests accurate estimation. Furthermore, the CI coverages for the regression parameters and disease risk are close to 0.95, which indicate Model 1 is able to accurately quantify uncertainty.

In order to compare the results from the model proposed in the previous chapter (Model 3) and this chapter (Models 4 and 5), I consider all four metrics for the regression parameter (β_1, β_2) and disease risk $R(\mathcal{G}_j)$. Firstly, let us consider the regression parameter β_1, which corresponds to the estimated covariate $\mathrm{x}_1(\mathcal{G}_j)$ disaggregated to the grid level. The results are presented in Table 5.2. I found that Models 3, 4 and 5 produce close to unbiased estimates for β_1. Note that some datasets for Model 4 are not included in the results since they produce non-converged MCMC chains, especially in scenarios 2, 4, 6, and 8 where I assumed 5% of population in each grid square have the disease event ($\psi = 0.05$). Thus for a prevalent disease it seems that data augmentation with τ^2 estimated in the model does not produce good results, as the variance τ^2 is over estimated, leading to poor inference. In terms of RMSE, Model 5 performs similar to Model 3. Furthermore, the CI coverages of Models 3 and 5 show no consistent pattern.

Next, consider regression parameter β_2, which corresponds to the true known grid level covariate $\mathrm{x}_2(\mathcal{G}_j)$. Models 4 and 5 produced unbiased estimates, while Model 3 produced biased estimates, which suggest the data augmentation approach is superior in this regard. Model 5 produced similar RMSEs to Model 4 but smaller than Model 3 by the factors between 3 and 7, therefore data augmentation is outperform in terms of RMSE. Furthermore, CI coverages for Model 3 are zero across all scenarios, this is because Model 3 produce biased estimates of the regression parameter β_2, and therefore the 95% CIs of the estimates do not include the true value of β_2.

Finally, Models 3, 4, and 5 produce close to unbiased estimates for the disease risk at the grid level $R(\mathcal{G}_j)$, as shown in Table 5.4, with RMSE values that are similar. Model 4 is the worst model for estimating the disease risk across all metrics, again due to overestimation of τ^2. Furthermore, it produces unstable results in some scenarios with biased estimates and extremely high RMSE values, therefore it is needed extra simulated datasets in order to obtain 100 converged datasets. Model 5 performs less well than Model 3 in terms of the CI coverage, however the average width of CI for Model 3 is approximately two times bigger than Model 5 across all scenarios. In conclusion, when consider all metrics for the regression parameters and disease risk at the grid level (grid of size 1,000 metres), I found that Model 5 performs the best, followed by Model 3. From the overall results, Model 4 produced awful estimates of disease risk at the grid level especially for a prevalence disease, therefore Model 4 will be removed from the simulation study for the grid squares with sides of length 500 metres.

Table 5.5: Results from the simulation study for the regression parameter $\beta_1 = 0.1$ with the estimated covariate at the grid level $x_1(\mathcal{G}_j)$ (grid size 500 m).

Metric	Scenario	Model 1	Model 3	Model 5
Bias	1	0.0018	0.1551	0.2469
	2	-0.0011	0.1482	0.2802
	3	0.0019	0.1737	0.3056
	4	0.0004	0.1655	0.2877
	5	0.0014	0.1725	0.3096
	6	0.0009	0.1670	0.2842
	7	-0.0007	0.1513	0.2633
	8	-0.0008	0.1477	0.2581
RMSE	1	0.0090	0.1730	0.2922
	2	0.0044	0.1583	0.2954
	3	0.0085	0.1895	0.3418
	4	0.0046	0.1728	0.3027
	5	0.0091	0.2010	0.3722
	6	0.0057	0.1801	0.3095
	7	0.0101	0.1767	0.3047
	8	0.0055	0.1650	0.2780
CI coverage	1	0.96	0.65	0.45
	2	0.95	0.15	0.02
	3	0.98	0.52	0.21
	4	0.96	0.03	0.02
	5	0.99	0.58	0.28
	6	0.95	0.05	0.02
	7	0.96	0.62	0.35
	8	1.00	0.19	0.03
Average width	1	0.0378	0.3657	0.4174
	2	0.0181	0.1726	0.1858
	3	0.0375	0.3694	0.4294
	4	0.0182	0.1686	0.1885
	5	0.0401	0.3908	0.4281
	6	0.0220	0.1956	0.1951
	7	0.0402	0.3542	0.4071
	8	0.0233	0.1778	0.1856

Table 5.6: Results from the simulation study for the regression parameter $\beta_2 = 0.1$ with the true covariate at the grid level $x_2(\mathcal{G}_j)$ (grid size 500 m).

Metric	Scenario	Model 1	Model 3	Model 5
Bias	1	0.0005	-0.0860	0.0039
	2	0.0003	-0.0881	-0.0022
	3	0.0014	-0.0867	-0.0086
	4	0.0004	-0.0866	-0.0002
	5	0.0008	-0.0873	0.0007
	6	-0.0009	-0.0889	-0.0015
	7	0.0011	-0.0861	-0.0001
	8	0.0015	-0.0876	0.0023
RMSE	1	0.0095	0.0861	0.0321
	2	0.0048	0.0882	0.0200
	3	0.0098	0.0867	0.0295
	4	0.0045	0.0867	0.0195
	5	0.0111	0.0874	0.0350
	6	0.0055	0.0890	0.0281
	7	0.0093	0.0862	0.0356
	8	0.0053	0.0877	0.0230
CI coverage	1	0.95	0	0.83
	2	0.93	0	0.71
	3	0.92	0	0.81
	4	0.97	0	0.67
	5	0.91	0	0.74
	6	0.96	0	0.37
	7	0.96	0	0.77
	8	0.97	0	0.55
Average width	1	0.0380	0.0370	0.0786
	2	0.0181	0.0169	0.0353
	3	0.0377	0.0371	0.0778
	4	0.0182	0.0167	0.0349
	5	0.0403	0.0376	0.0783
	6	0.0220	0.0173	0.0349
	7	0.0403	0.0370	0.0770
	8	0.0234	0.0170	0.0351

Table 5.7: Results from the simulation study for the disease risk at the grid level $R(\mathcal{G}_j)$ (grid size 500 m).

Metric	Scenario	Model 1	Model 3	Model 5
Bias	1	0.0005	-0.0017	0.0112
	2	-0.0002	-0.0045	0.0043
	3	0.0001	-0.0005	0.0112
	4	0.0001	-0.0009	0.0022
	5	0.0000	-0.0109	0.0133
	6	-0.0004	-0.0074	0.0018
	7	0.0010	0.0006	0.0118
	8	0.0007	-0.0021	0.0043
RMSE	1	0.0914	0.1615	0.2948
	2	0.0797	0.1591	0.1903
	3	0.0733	0.1514	0.6516
	4	0.0693	0.1508	0.2133
	5	0.1879	0.2439	1.1267
	6	0.1379	0.2078	0.2319
	7	0.1568	0.2110	0.5147
	8	0.1406	0.2095	0.3271
CI coverage	1	0.94	0.44	0.18
	2	0.95	0.39	0.13
	3	0.92	0.36	0.18
	4	0.94	0.28	0.10
	5	0.94	0.53	0.20
	6	0.95	0.43	0.14
	7	0.93	0.29	0.14
	8	0.95	0.33	0.10
Average width	1	0.3428	0.1874	0.0987
	2	0.2980	0.1653	0.0574
	3	0.2682	0.1408	0.1287
	4	0.2653	0.1093	0.0484
	5	0.6574	0.3401	0.1639
	6	0.5235	0.2361	0.0771
	7	0.5763	0.1557	0.1180
	8	0.5380	0.1773	0.0618

I repeat the simulation study with grid squares of size 500 metres to check whether the size of grid square affects the general pattern of results. Since Model 4 performed poorly for grid squares with side of length 1,000 metres, I removed it from this simulation study. The results from all metrics and models are presented in Tables 5.5, 5.6 and 5.7. Model 3 is better than Model 5 for the estimated grid level covariate, $\mathrm{x}_1(\mathcal{G}_j)$, and the disease risk, $R(\mathcal{G}_j)$ across all metrics and scenarios. Model 5 is better than Model 3 for the true grid level covariate, $\mathrm{x}_2(\mathcal{G}_j)$, since it produced unbiased estimates with smaller RMSE values and higher CI coverages.

The key findings of these results are broadly similar to those for grid squares of size 1,000 metres. However, the results from the models fitted to data with grid squares of size 500 metres are worse than grid squares of size 1,000 metres for most metrics and scenarios. For example, Model 5 produced biased estimates for the regression parameter related to the estimated covariate $[\mathrm{x}_1(\mathcal{G}_j)]$, while it produced unbiased estimates for grid squares of size 1,000 metres. Furthermore, in terms of RMSE in regression parameters and disease risk across all scenarios, Models 3 and 5 have higher RMSE than the results from grid squares of size 1,000 metres. This is because I have the same amount of actual data at the areal unit level and then I use these data to estimate the disease risk and model parameters at the grid level. The number of grid squares of size 500 metres is greater than the number of grid squares of size 1,000 metres, which means that I have to estimate more grid level parameters using the same amount of data, which leads to less accurate estimation.

5.4 Application to real data

This section continues the analysis of the respiratory hospitalisation data presented in Chapters 3 and 4.

5.4.1 Data description

As in the previous chapters, the study region is the Greater Glasgow and Clyde Health Board area, and I use the respiratory hospital admission data presented in Section 3.2. The response disease data $\boldsymbol{Y}(\mathcal{A}) = [Y(\mathcal{A}_1), ..., Y(\mathcal{A}_n)]$ are based on the period from

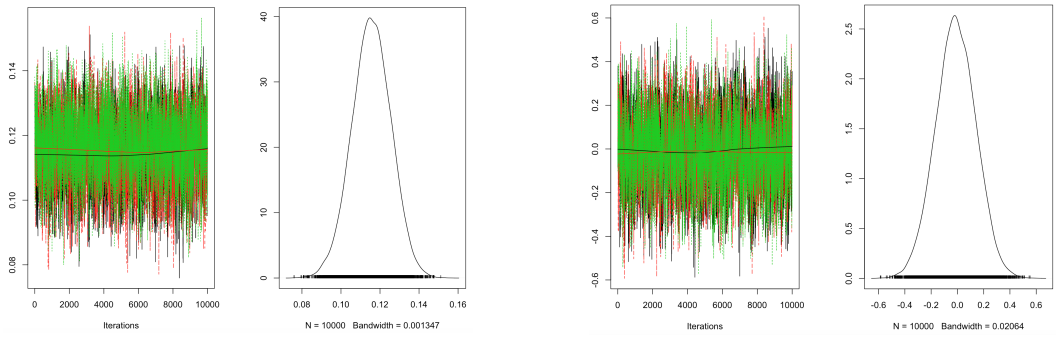
January 2015 to December 2016, where $Y(\mathcal{A}_i)$ is the number of hospital admissions with a primary diagnosis of respiratory disease in intermediate zone i. The expected values, $\boldsymbol{e}(\mathcal{A}) = [e(\mathcal{A}_1), ..., e(\mathcal{A}_n)]$ are the expected hospital admission numbers for each areal unit, which are computed via indirect standardisation as described in Section 2.3.1.

5.4.2 Results

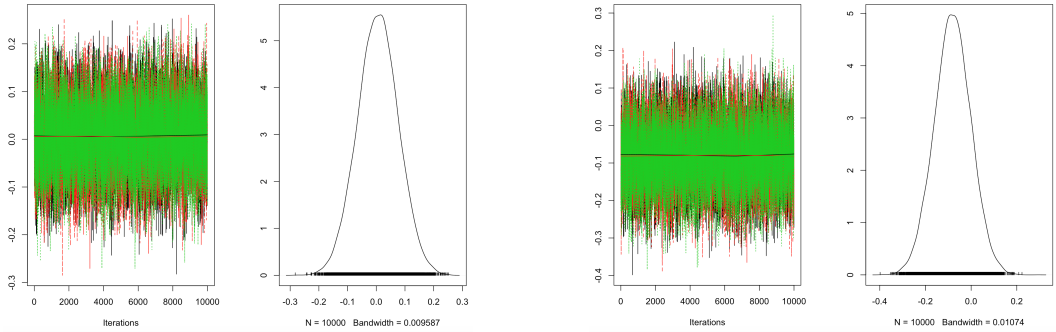
In order to compare the performance of Model 3 from the previous chapter and the proposed model (Model 5), Model 5 has been fitted to the respiratory disease data with the grid square sides of lengths of 1,000 and 500 metres. The results are obtained by using MCMC inference, based on 200,000 samples with 50,000 burn-in samples and thinned by a factor of 15, leaving 10,000 samples for model inference. Note that the sensitivity analysis is not carried out since the variance parameter of random effect (τ^2) is fixed for Model 5.

Convergence diagnostic

The method of Gelman and Rubin (Gelman and Rubin, 1992) and trace plots assessment are used to diagnose the convergence of the MCMC chains. Since there are a large number of parameters in the model especially for the 500-metre grid size, therefore it is infeasible to check all of them. In order to do this, ten parameters are randomly checked for convergence of the chain and only selected parameters are presented. Figures 5.1 to 5.2 illustrate the plots of each model parameter from the model proposed in this chapter (Model 5) with the grid square of sizes for both 1,000 and 500 meters. While the diagnostic for Model 3 (multiple imputation) which is used as a comparative model does not present here since it is similar to the previous chapter. The figures show that there is no clear pattern appear in all selected parameters, which indicate that the chains appear to have converged. In addition, the Gelman-Rubin diagnostic used to diagnose convergence for multiple chains, the result shows that the Gelman-Rubin values are less than 1.1. It is suggested that the posterior distributions are well mixed.

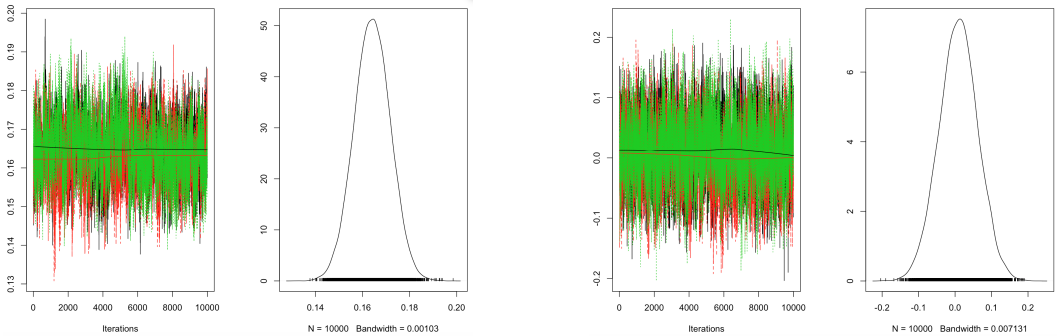


(a) β (b) $\phi(\mathcal{G}_1)$

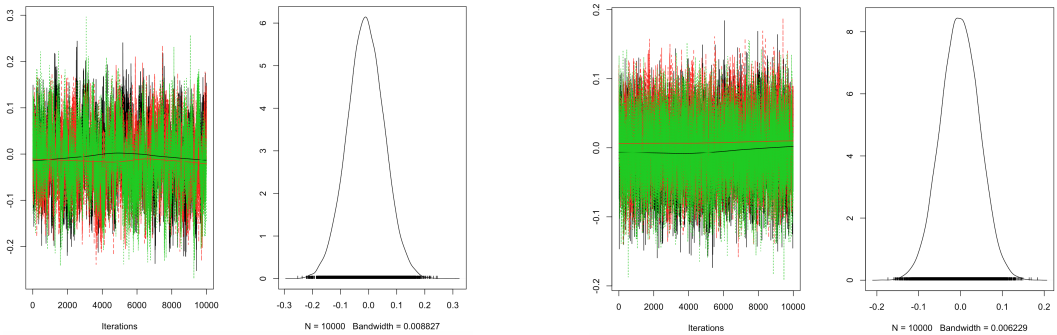


(c) $\phi(\mathcal{G}_{75})$ (d) $\phi(\mathcal{G}_{289})$

Figure 5.1: Traceplots of MCMC samples for selected parameter from Model 5 (grid of size 1,000m).



(a) β (b) τ^2



(c) ρ (d) $\phi(\mathcal{G}_1)$

Figure 5.2: Traceplots of MCMC samples for each parameter from Model 2 (grid of size 500m).

Posterior predictive check

The predictive posterior model check is carried out in order to check the appropriateness of the model for the data. Here, the observed data at the grid level are unknown, therefore the observed data at the areal unit level have been applied to compare to the simulated data (grid level) from the fitted model that aggregates to the areal unit level. Figure 5.3 shows the relationship between the observed disease counts at the areal unit level and the simulated data for the grid square sizes of 1,000 and 500 metres. The plots show that these two datasets are fairly similar since these data lie on the straight line, however there are some outliers. This is because the simulated data are transformed twice (disaggregate to the grid level and aggregate back to the areal unit level), therefore some information might be lost during the process. Hence, these results indicate that the models fit the data well and appropriately to make a model inference.

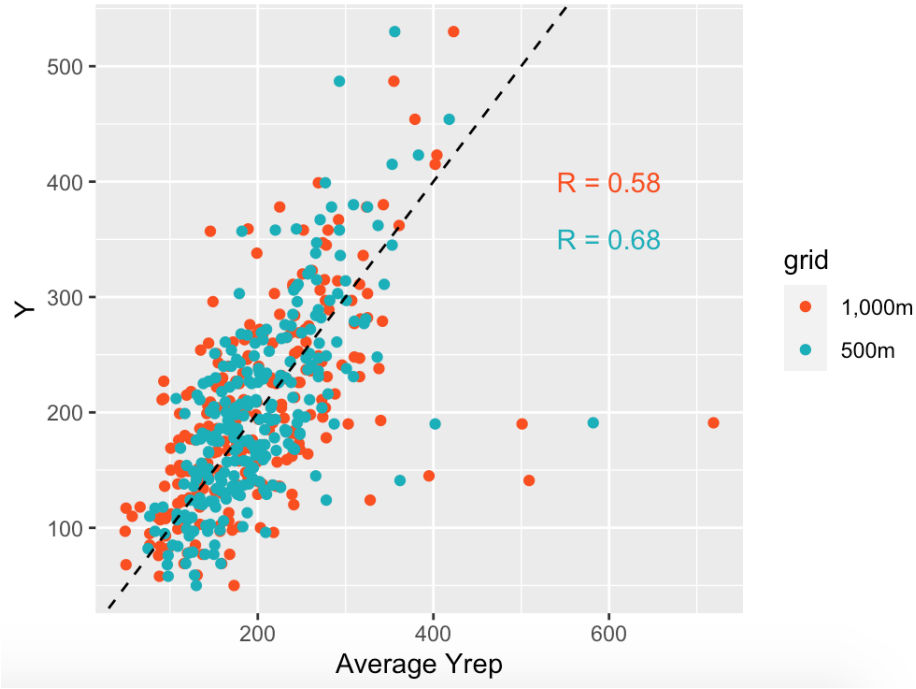


Figure 5.3: Posterior predictive checks (Model 5)

Main results

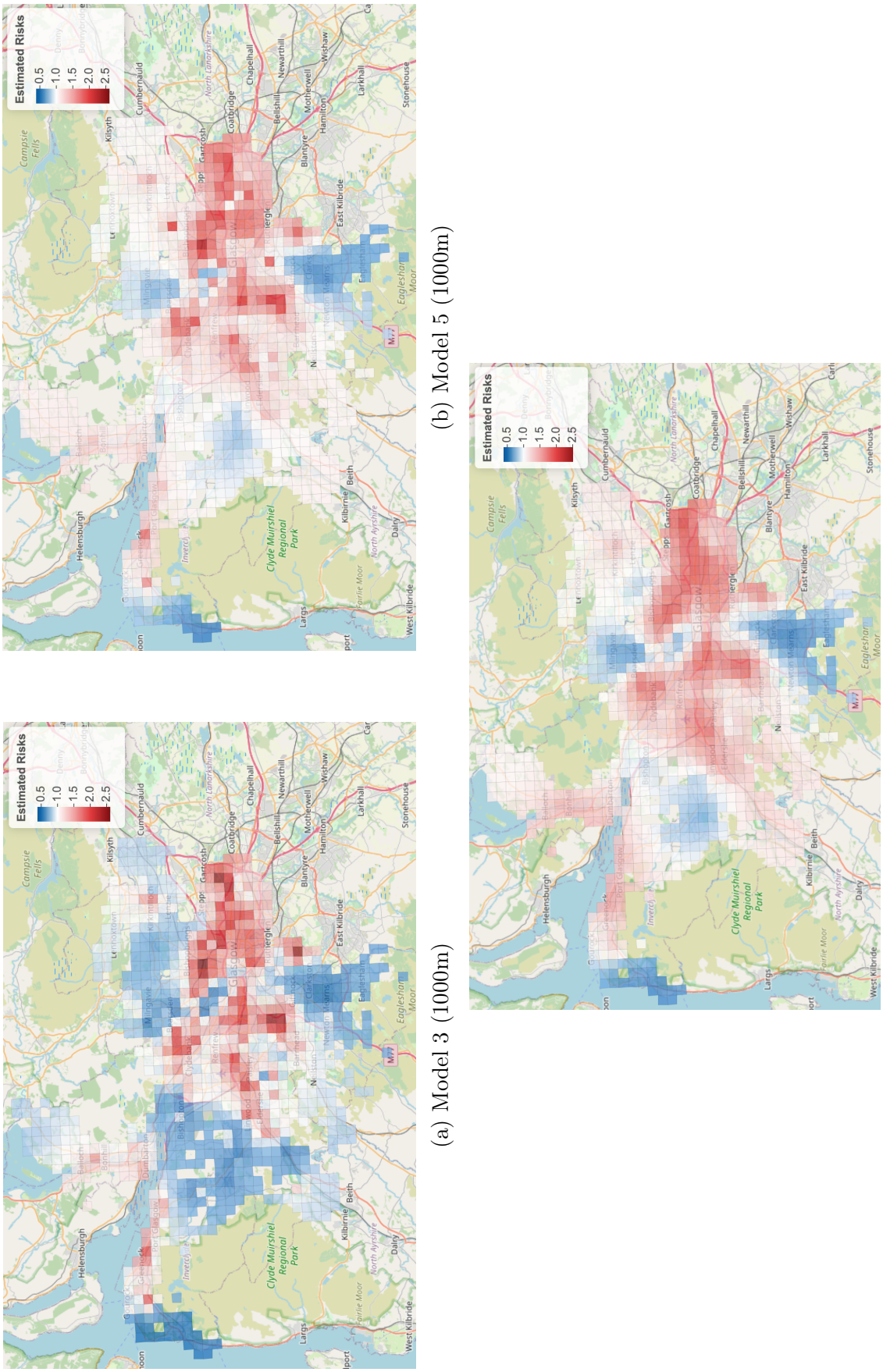
Figures 5.4 and 5.5 display the estimated disease risk from Models 3 and 5 with the grid squares of sizes of 1,000 and 500 metres. Model 3 is the model proposed in the previous

chapter (multiple imputation), while Model 5 is the proposed model in this chapter (data augmentation with fixed τ^2). Here $\hat{\tau}^2$ is estimated by (5.2.13) and $\hat{\tau^2} = 0.0158$ for grid squares of size 1,000 metres and 0.0044 for 500 metres . Overall, Model 5 produces similar disease maps for both grid squares of sizes 1,000 and 500 metres, as can be seen in Figures 5.4b and 5.5b. The regions with the higher disease risks tend to be in the east (e.g. Easterhouse, Shettleston), and the north of Glasgow city centre (e.g. Possilpark, Springburn), as well as the south-west (e.g. Priesthill, Govan) and the north-west (e.g. Clydebank, Drumchapel). On the other hand, the regions of lower risk tend to be in the south-west (e.g. Whitecraigs and Newton Mearns), and the West End of the city centre (e.g. Kelvinside and Jordanhill). These results are similar to the results of Model 3 which are presented in Figures 5.4a and 5.5a. As in the previous chapter, I note that people in areas that are less wealthy are more likely to be hospitalised for respiratory disease than those in more wealthy areas.

Figures 5.4 and 5.5 show that Model 3 and Model 5 produce the similar pattern of disease maps. However, Model 3 has more spatial variation than Model 5 because Model 3 produces an estimated τ^2 which is greater than Model 5 with a similar estimated ρ. For example, the estimated τ^2 for grid squares of size 500 metres being: 0.0872 (Model 3) vs 0.0046 (Model 5), which in Model 3 is greater than Model 5 by a factor of 20. Furthermore, there are different estimates between the two models in some grid squares. To quantify the differences between the models, Figure 5.6 presents the correlation of the estimated disease risks from these models. For the grid squares of size 1,000 metres, the correlation between the estimates of Model 3 and Model 5 is 0.91. For the grid square of size 500 metres, the correlation between the estimates of Model 3 and Model 5 is 0.89. This unsurprisingly suggests a strong correlation between the models. Figure 5.7 shows the absolute estimated risk differences between Model 3 and Model 5 versus the average estimated risk for both grid squares of sizes 1,000 and 500 metres. These plots have a U-shaped pattern which suggest that these two models tend to agree when the disease risk is between 0.75 and 1.75 but tend to disagree more strongly outside this range.

In order to select the better model, the standard Bayesian model selection criteria such as Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is inappropriate in this study for a few reasons. In the multiple imputation approach, the results are based on a combination of ten different datasets. While DIC is appropriate to compare different models in the same dataset. Furthermore, the grid level disease counts are updated in every 100th iterations for data augmentation and hence they are not a single dataset. Consequently, in order to carry out the model selection, the sensible method is to transform the estimate disease counts at the grid level from each model into the areal unit level $[\hat{Y}(\mathcal{A}_i)]$ via 4.3.3, and then compare them to the observed disease counts $[Y(\mathcal{A}_i)]$ using $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[Y(\mathcal{A}_i) - \hat{Y}(\mathcal{A}_i)]^2}$, where n is the number of areal units.

Here I aim to compare the proposed models in this study to an existing model in order to select the best model. The model proposed by Taylor et al. (2018) is initially considered as a competitive model but there is an issue with software package that could not be applied to the real data in this study. Therefore it is proposed for the simple method: Kriging (Krige, 1951), a geostatistical process aims to predict unobserved data at unobserved locations. Since the main aim of this study is to estimate disease risks at the gird level, therefore Kriging is used as a competitive model for selecting the best model. The detail of Kriging is presented in Section 2.3.5. Moreover, the SIR values for each IZ are used to estimate the disease risk at the grid level by using (2.3.14). The results show that these three models produce the similar patterns of estimated disease risks at the grid level, as can be seen in Figures 5.4 and 5.5. However Model 3 (multiple imputation) produces smallest RMSE followed by Model 5 (data augmentation) and the method of Kriging as presented in Table 5.8, this is true for both sizes of the grid square. Therefore, the best approach to estimate disease risks at the grid level is multiple imputation which corresponds to the results from the simulation study. These results are not surprising since Kriging assumes that the mean and variance of the SIRs are constant across the study region (stationarity assumption), which is not met in this application. In addition, Kriging is likely to perform badly because it assumes that the areal unit SIR is represented by its centroid at a single point.

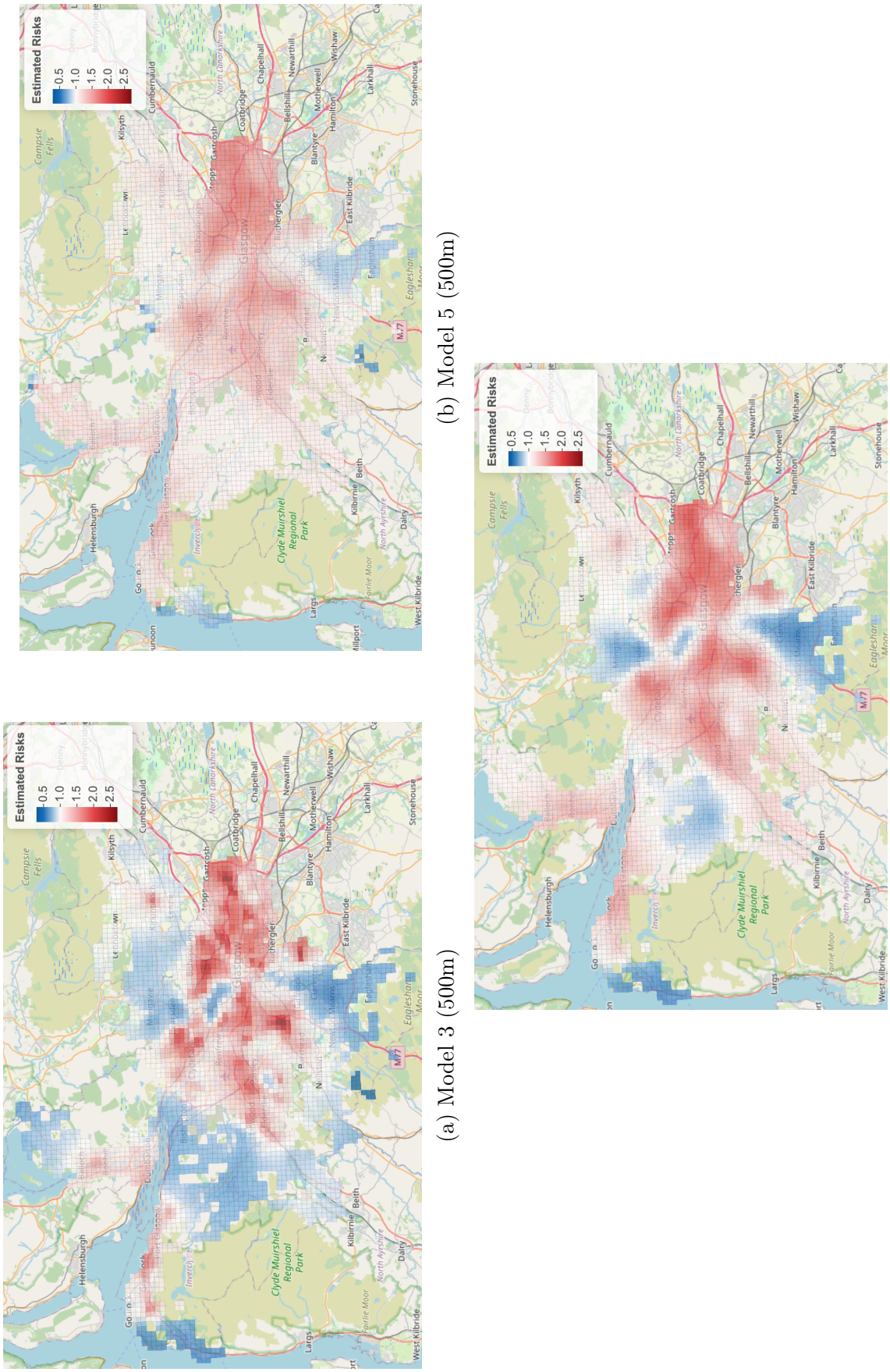


(a) Model 3 (1000m)

(b) Model 5 (1000m)

(c) Kriging (1000m)

Figure 5.4: Estimated disease risks from the proposed models on grid square size 1,000 metres.



(a) Model 3 (500m)

(b) Model 5 (500m)

(c) Kriging (500m)

Figure 5.5: Estimated disease risks from the proposed models on grid square size 500 metres.

Table 5.8: RMSE values of disease counts at the areal unit level.

Model	RMSE	
	1,000 m	500 m
Model 3	62.99	39.67
Model 5	73.82	56.71
Kriging	77.73	61.61

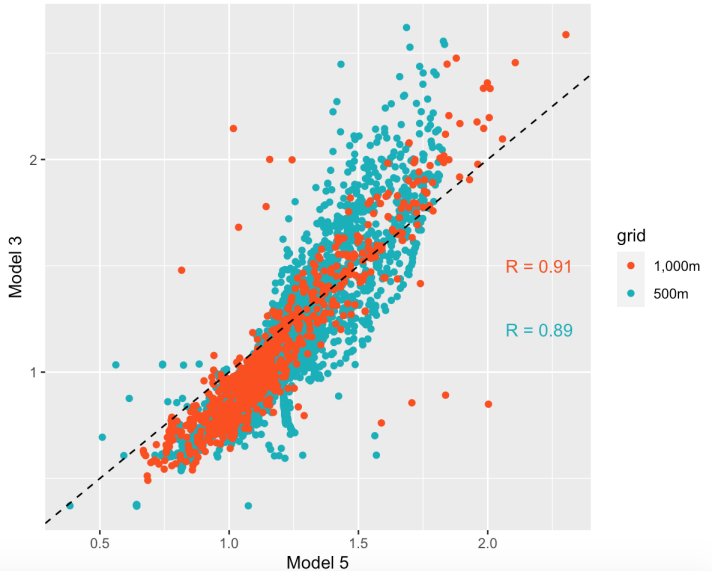


Figure 5.6: Correlation between the estimated disease risk of Models 3 and 5.

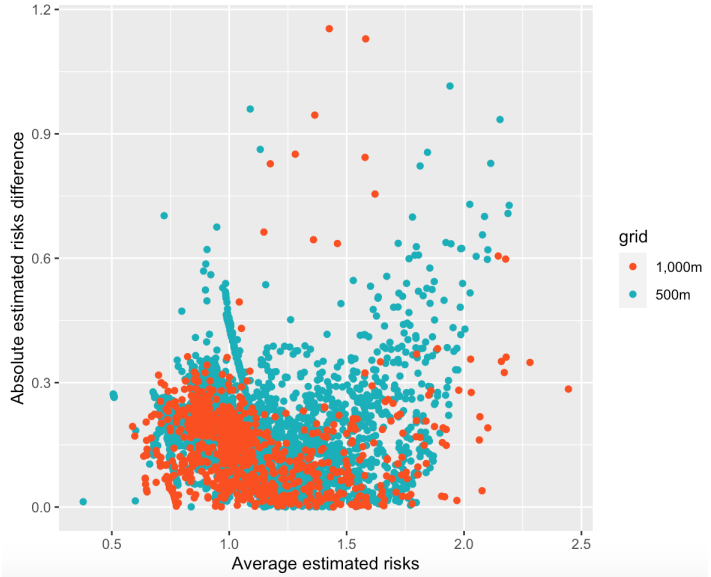
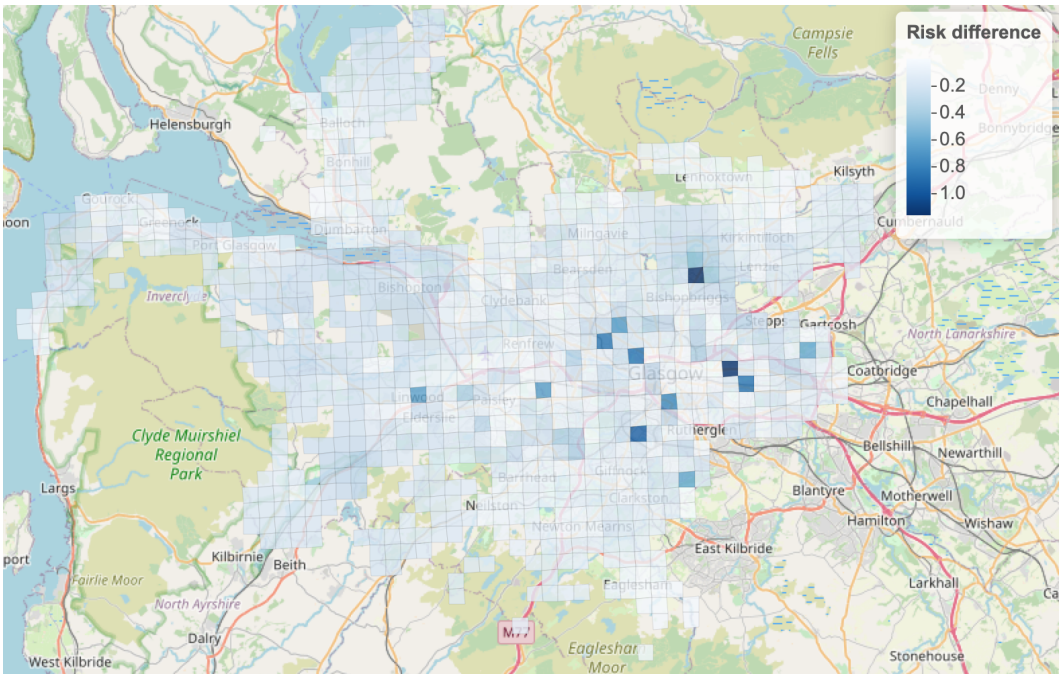
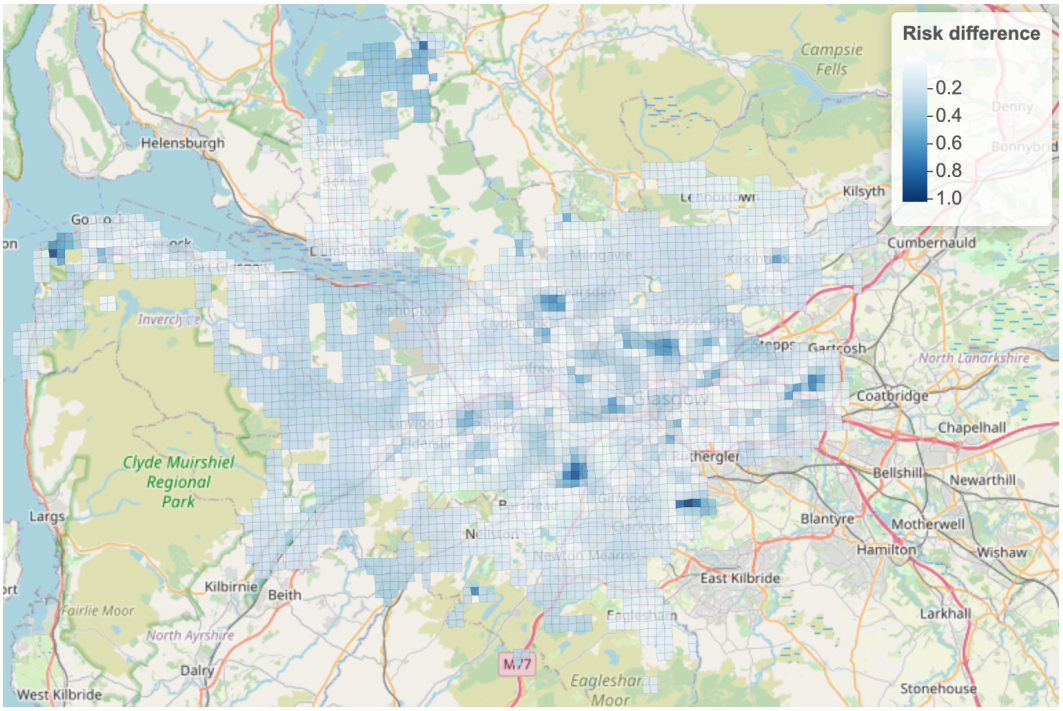


Figure 5.7: Plots of the absolute estimated disease risk difference between Models 3 and 5 versus the average of the estimated disease risk.



(a) Grid size 1,000 metres



(b) Grid size 500 metres

Figure 5.8: The estimated disease risk difference between Models 3 and 5.

5.5 Conclusion

Here I have proposed statistical methodology which estimates the spatial pattern in disease risk at the grid level via a data augmentation approach. This method allows for uncertainty in the grid level disease counts when estimating grid level disease risk, unlike the multiple imputation approach in the previous chapter. A Leroux CAR model was initially used to estimate disease risk but this model contains a spatial smoothness parameter, ρ, which allows the spatial autocorrelation in the data to range from very weak to very strong. Exploratory analyses found this could lead to high variation in the estimates of disease risk and disease cases at the grid level,which could lead to unrealistic estimation. Therefore I enforce strong spatial autocorrelation of the disease risk by setting $\rho = 1$, which corresponds to the intrinsic CAR model proposed by Besag et al. (1991).

A simulation study was conducted to examine how well the proposed models can estimate disease risk and model parameters at the grid level. In this simulation study, I compared four models across two different grid sizes (1,000 and 500 metres). Model 1 was fitted to the true grid level data, which was used as a reference model since it was expected to perform the best. Model 3 was the proposed model in the previous chapter (multiple imputation). Model 4 and Model 5 were both data augmentation approaches which were introduced in this chapter. In Model 4 the spatial random effect variance, τ^2, was estimated as part of the MCMC algorithm, while in Model 5 τ^2 was estimated by empirical Bayes methods prior to fitting the MCMC algorithm and was held constant throughout.

The key results for grid squares sides of lengths 1,000 and 500 metres are roughly similar. Model 1 is the best model across all scenarios and metrics, which is expected since it was fitted to the true grid level data. Model 4 performed worst in terms of estimating disease risk at the grid level due to τ^2 being overestimated in some simulated datasets, which leaded to extreme disease risk estimates at the grid level. This is likely because of a lack of identifiability from estimating disease counts and model parameters together. Model 4 is therefore not recommended; it is clear that a data augmentation approach

needs to use Model 5, where τ^2 is estimated by empirical Bayes methods.

Model 3 and Model 5 generally produced similar results except for the regression parameter related to the true known grid level covariate, where Model 5 performed better than Model 3. This is due to Model 3 producing biased estimates, which also leads to poorer credible interval coverage across all metrics and scenarios. This is likely because the true grid level covariate has not been transformed to the grid level via an aggregation and disaggregation processes in the way that other variables have, and consequently it has a different scale of spatial smoothness to the other variables. This causes Model 3 to consistently underestimate the covariate effect, leading to biased results. Model 3 is therefore not recommended for estimating the regression parameter related to the true known grid level covariate; Model 5 must be used in such cases.

The overall pattern of the results from the models fitted to the data with grid squares sides of length 500 metres is mainly similar to those with grid squares sides of length 1,000 metres. However, the RMSEs across all models and parameters for grid squares of size 1,000 metres are smaller than grid squares of size 500 metres by a factor of between 1.3 and 12, which suggests that grid squares of size 1,000 metres produce more precise estimation than grid squares of size 500 metres. This is because we have the same number of data points regardless of how many grid squares we have. Therefore to estimate on the 1,000 metres scale we need to estimate 853 grid square parameters from 257 data points, while to estimate on the 500 metres scale we need to estimate 3,106 grid square parameters from the same 257 data points. It therefore follows that estimation on the smaller grid squares leads to less accuracy. However, I note from Figures 5.4 and 5.5 that as the size of the grid square increases, the disease risk map can become less smooth and begins to look somewhat pixellated. This is a trade off between accuracy of the estimation and smoother mapping visualisation, as if we make the grid squares too large they might cover areas with significantly different disease risk which could lead to less useful estimates. On the other hand if the grid squares are too small, we will estimate too many grid squares too close together which will have the same risk level. We would also need to estimate too many parameters which could lead to less precise estimates as shown in the simulation study.

Finally, I compare the performance of the reference model (Model 1) to the proposed models in the previous chapter (Model 3) and this chapter (Model 5). Model 3 and Model 5 produce higher RMSE values than Model 1 across all scenarios and parameters by a factor of between 1.8 and 4. Although Model 3 and Model 5 perform less well than Model 1, they have adequate ability in terms of estimating disease risk and model parameters at the grid level, which is expected from these proposed models.

Chapter 6

Spatio-temporal modelling of respiratory disease in Glasgow

6.1 Introduction

The term "health inequalities" generally refers to differences in health status or in the distribution of health determinants between individuals or population groups (World Health Organization, 2013). For example, differences in morbidity between men and women or between people from different social classes. There is sufficient evidence that education, employment status, and income are influential factors on people's health (Mackenbach et al., 2008). These unfair and avoidable differences in people's health exist both within and between countries. For example, there are nearly 20% of children in the UK who live in poverty areas that have much worse health outcomes than those live in wealthier areas since they face multiple risks for future poor health e.g. diet and lifestyle choices (The Lancet Respiratory Medicine, 2017). Furthermore, people in low socioeconomic status countries exhibit lower lung function and more respiratory symptoms than those in higher socioeconomic status countries (Pleasants et al., 2016).

The first major publication about health inequality in the UK was the Black report by Black et al. (1982). The report showed that ill-health and death are unequally distributed across the UK, and suggested that these inequalities were attributable to socio-economic inequalities affecting health. This finding was confirmed by several

studies e.g. Acheson (1998), Marmot et al. (2010), Bartley (2016). In addition, the social class differences in mortality rates had widened within ten years of the Black report (Smith et al., 1990). Life expectancy is widely used as a proxy indicator of people's health. Ellis and Fry (2010) argued that life expectancy in the North East of England is consistency lower than other regions, and female life expectancy is higher than males in all regions. Levin and Leyland (2006) compared health inequalities in urban and rural areas in Scotland. They found that health inequalities in Scotland have grown over time, and people in remote rural areas have witnessed a significant rise in inequality, especially those aged over 65 years. A more recent study by Jack et al. (2019) proposed a multivariate spatiotemporal model for estimating small area variation in disease risk, and they concluded that health inequalities in cerebrovascular and coronary heart disease were reducing over time, but inequalities in respiratory disease appear to be growing worse over time (2003 - 2012).

In this chapter, I focus on measuring health inequalities in the Greater Glasgow and Clyde Health Board area. This is because the life expectancy for men in Glasgow is the lowest of any major city in the UK, as presented in Figure 6.1 taken from Walsh et al. (2016). Although the trend for Glasgow is slightly increasing, the gaps between Glasgow and other cities are slightly widening. Additionally, life expectancies between people in different areas in Glasgow are very different (NHS Health Scotland, 2015), as is illustrated in Figure 6.2. Life expectancy for men living in Bridgeton, a less wealthy area, is on average 14.3 years lower than for those living in Jordanhill, a wealthier area, even though they live only seven stations apart from each other (15 minutes of travelling). Similarly, this difference in life expectancy for women is 11.7 years. It is therefore clear that Glasgow has large health inequalities, and it is of interest to investigate these in more detail.

There were several previous research studies on health inequality in Glasgow, most of them focused on either health inequality between intermediate zones within Glasgow (Hanlon et al., 2006; McCartney, 2010) or compared Glasgow to other cities in the UK Shelton (2009); Walsh et al. (2010, 2016). However, as described in previous chapters, an IZ level analysis assumes that disease risk is constant within an IZ, which is not nec-

essarily realistic. Therefore I extend the grid square level pseudo-continuous approach presented in the previous chapters to the spatio-temporal domain, thus allowing risk to vary pseudo-continuously over space. This chapter has two main aims. The first is to estimate the spatio-temporal variation in disease risk at the grid square level, by extending the methodology developed in previous chapters. The second is to compare the results obtained at the two grid square scales (sides of lengths 1,000 and 500 metres) used in this thesis, with the commonly used approach of undertaking inference at the areal unit IZ level. In undertaking the data analysis, the main motivating objectives are to answer three questions of interest as follows:

i) What is the average trend over time of respiratory disease risk across the Greater Glasgow and Clyde Health Board area?

ii) How has the respiratory disease risk in each part of Glasgow changed over time in the Greater Glasgow and Clyde Health Board area from 2013 - 2016?

iii) How are the health inequalities changing over time in the Greater Glasgow and Clyde Health Board area for respiratory disease risk?

I consider these three questions because I would like to estimate the trend in respiratory disease, whether it is increasing, decreasing, or stable, both in the whole Glasgow health board area and by locality within the study region. Understanding these trends would allow the health board to make a localised intervention in the part of Glasgow with an increased trend, and also to investigate the factors that influence these trends. Furthermore, estimating the changing level of health equality over time can provide information as to whether their inequality reduction plans are working.

The remainder of this chapter is organised as follows. Section 6.2 outlines the data being used in this chapter. Section 6.3 outlines the methods used to transform the areal unit data into grid level data. Then Section 6.4 presents the spatio-temporal model used in this study, and the results obtained from the model are shown in Section 6.5. Finally Section 6.6 summaries the main finding from the model fitting and discusses the advantages and disadvantages of this study.

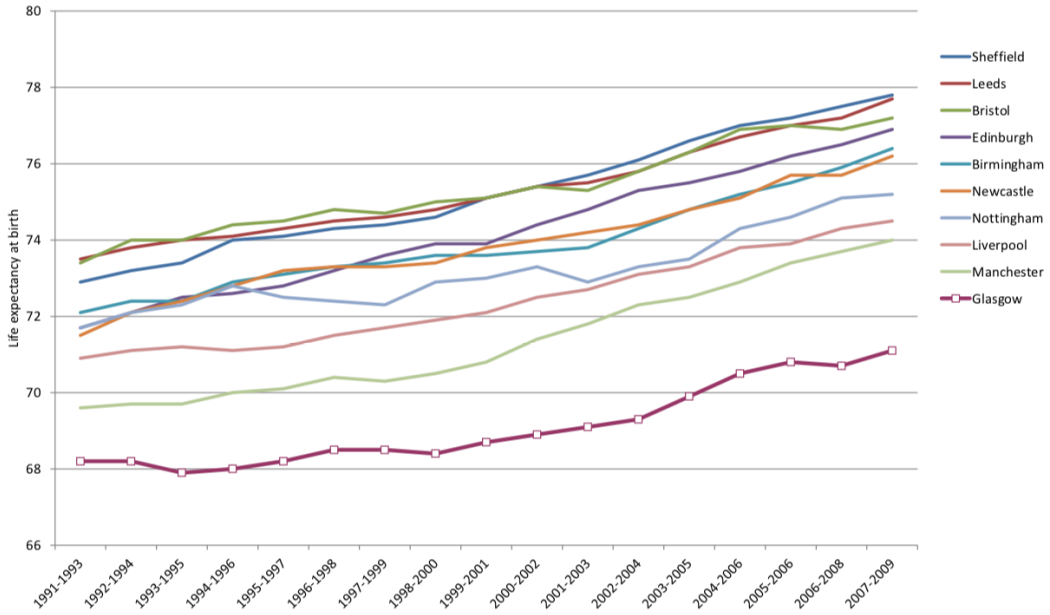


Figure 6.1: Male life expectancy for Glasgow compared with other UK cities, 1991-93 to 2007 - 09 (Walsh et al., 2016).

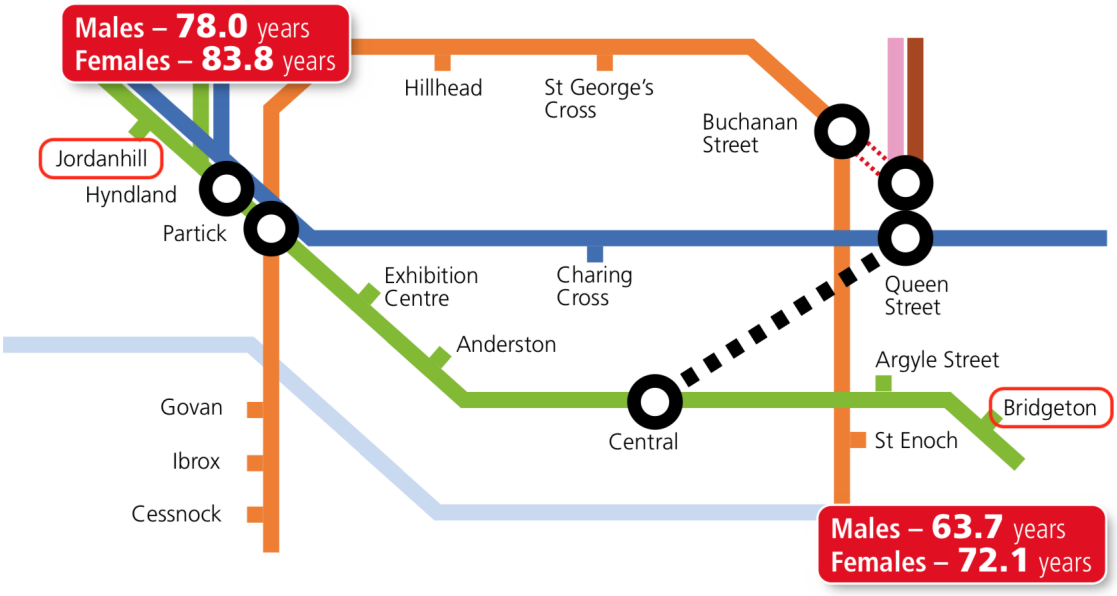


Figure 6.2: Part of the train map of Glasgow with life expectancy (NHS Health Scotland, 2015).

6.2 Data

The study region is the Greater Glasgow and Clyde Health Board area, which is the same area as studied in previous chapters. The disease data are yearly counts of the number of hospital admissions with a primary diagnosis of respiratory disease for the years 2013 to 2016 in each intermediate zone (IZ). Here I use four years of data because they are the most recent data that I can obtain from NHS Scotland which have the same spatial scale (IZ boundaries). The observed disease counts, $\boldsymbol{Y}_t(\mathcal{A}) = [Y_t(\mathcal{A}_1), ..., Y_t(\mathcal{A}_n)]$, represent the number of hospital admissions due to respiratory disease, where $Y_t(\mathcal{A}_i)$ denotes the disease count in region $\mathcal{A}_i$ in year t, where $i = 1, ..., n$ and $t = 1, ..., N$. The expected disease counts, $\boldsymbol{e}_t^*(\mathcal{A}) = [e_t^*(\mathcal{A}_1), ..., e_t^*(\mathcal{A}_n)]$, are the expected hospital admission numbers for each region and year, which are calculated separately for each year via indirect standardisation based on age and sex adjusted rates for the whole of Scotland. However, $e_t^*(\mathcal{A}_i)$ should be averaged over the entire period of time in order to explore the overall change in SIR over time across the entire region. The formula of averaging $e_t^*(\mathcal{A}_i)$ is as follows;

$$e_t(\mathcal{A}_i) = \frac{\sum_{t=1}^{T} e_t^*(\mathcal{A}_i)}{T}. \tag{6.2.1}$$

Figure 6.3 shows bloxplots of the SIR in IZs between the years 2013 and 2016. It can be seen that the median of the SIRs are slightly increasing. Figure 6.4 illustrates the spatial map of the SIR for each IZ in the Greater Glasgow and Clyde Health Board in 2016. The areas with lower SIRs are mostly rural areas e.g. Milngavie, Bishopton and Eagleshame and also areas in the south of the city centre of Glasgow e.g. Clarkston and Newton Mearns. In contrast, the areas with higher SIRs are mostly located in the east of the map, the airport areas e.g. Paisley and also areas in the north-west e.g. Clydebank. The areas with lower SIR values tend to be the wealthy areas, but on the other hand the areas with higher SIR values tend to be the less wealthy areas. In other words, the affluent areas tend to have lower SIRs than the less affluent areas.

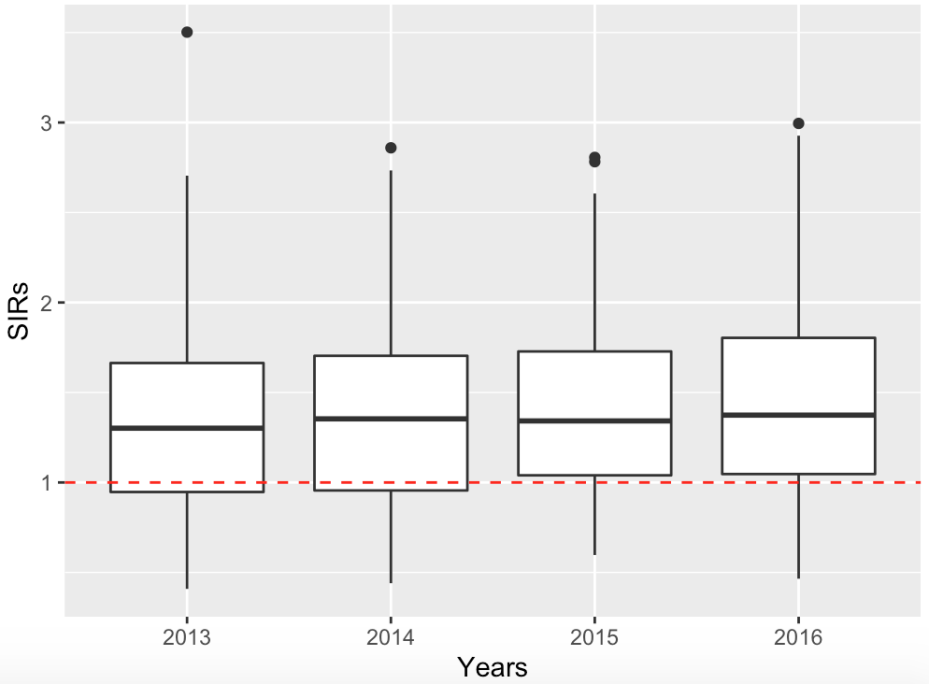


Figure 6.3: Boxplots of the of the standardised incidence ratio (SIR) for respiratory disease hospital admissions from 2013 to 2016.

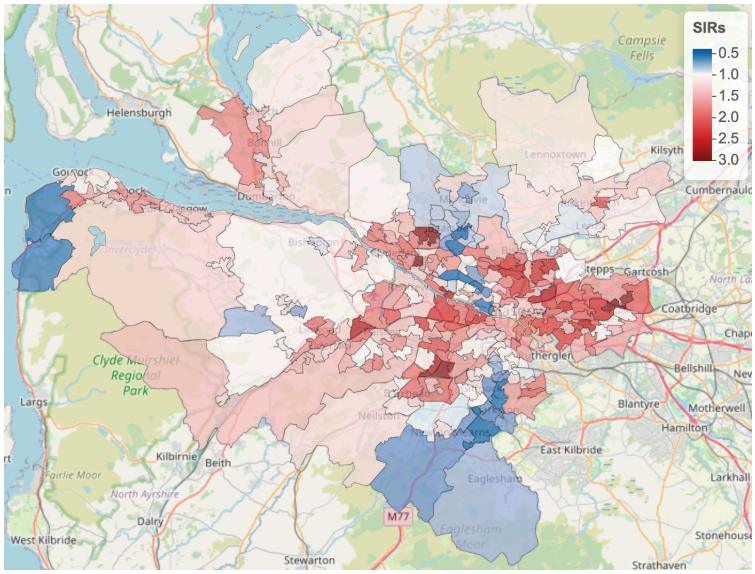


Figure 6.4: The standardised incidence ratio (SIR) for respiratory disease for each IZ in the Greater Glasgow and Clyde Health Board in 2016.

6.3 Methodology

As the aim of this chapter is to make grid square level inference based on areal unit level data, the observed and expected disease counts at the areal level $[Y_t(\mathcal{A}_i), e_t(\mathcal{A}_i)]$ have to be transformed to the grid level $[Y_t(\mathcal{G}_j), e_t(\mathcal{G}_j)]$ before being modelled to estimate the grid level pattern in disease risk. This is achieved in two stages, which are outlined in sections 6.3.1 and 6.3.2 below. The approach adopted for this transformation is similar to Chapter 4.

6.3.1 Grid level expected disease counts $[e_t(\mathcal{G}_j)]$

First of all, I must allocate the average expected disease counts at the areal unit level, $e_t(\mathcal{A}_i)$, to the m grid squares, $e_t(\mathcal{G}_j)$. When carrying out this process, one must ensure that the total numbers of expected counts at the areal unit level and the grid square level are the same, i.e. $\sum_{i=1}^{n} e_t(\mathcal{A}_i) = \sum_{j=1}^{m} e_t(\mathcal{G}_j)$. Letting $e_t(\mathcal{A}_i \cap \mathcal{G}_j)$ represent the expected count in the intersection area between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$ in year t, it is clear that the expected disease count in grid square $\mathcal{G}_j$ is the sum of expected disease counts in the intersection areas between grid square $\mathcal{G}_j$ and all regions $\mathcal{A}_i$, that is $e_t(\mathcal{G}_j) = \sum_{i=1}^{n} e_t(\mathcal{A}_i \cap \mathcal{G}_j)$. Assuming then the expected disease counts are distributed proportionally to the population density, an initial estimate of $e_t(\mathcal{G}_j)$ is

$$e_t(\mathcal{G}_j) = \sum_{i=1}^{n} e_t(\mathcal{A}_i \cap \mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{k=1}^{m} P(\mathcal{A}_i \cap \mathcal{G}_k)} e_t(\mathcal{A}_i), \tag{6.3.1}$$

where $P(\mathcal{A}_i \cap \mathcal{G}_j)$ is the population in the intersection area between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$, which is unknown. This unknown population size can be estimated based on multiplying the population size in grid square $\mathcal{G}_j$, $P(\mathcal{G}_j)$, by the proportion of that grid square that covers region $\mathcal{A}_i$. Grid level population data can be obtained from Reis et al. (2017). Therefore I estimate

$$P(\mathcal{A}_i \cap \mathcal{G}_j) = \frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)} P(\mathcal{G}_j). \tag{6.3.2}$$

This assumes that population density is constantly distributed across grid square $\mathcal{G}_j$. Finally, substitute $P(\mathcal{A}_i \cap \mathcal{G}_j)$ from (6.3.2) into (6.3.1), so that

$$e_t(\mathcal{G}_j) = \sum_{i=1}^{n} \frac{P(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} P(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}} e_t(\mathcal{A}_i). \tag{6.3.3}$$

It is straightforward to show that the sum of the grid level expected counts is equal to the sum of the areal unit level expected counts $\left(\sum_{i=1}^{n} e_t(\mathcal{A}_i) = \sum_{j=1}^{m} e_t(\mathcal{G}_j)\right)$. Note that, the expected disease counts at the grid level in each grid square are similar for every year.

6.3.2 Grid level disease counts $[Y_t(\mathcal{G}_j)]$

Disease counts in grid square $\mathcal{G}_j$ in year t, $Y_t(\mathcal{G}_j)$, is estimated via a multiple imputation approach which is similar to that described in Section 4.2.3. Then fit a spatio-temporal model to these data to estimate disease risk at the grid level. This general approach is implemented as follows.

Let us denote the disease count in the intersection area between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$ in year t by $Y_t(\mathcal{A}_i \cap \mathcal{G}_j)$. It is clear that

$$Y_t(\mathcal{G}_j) = \sum_{i=1}^{n} Y_t(\mathcal{A}_i \cap \mathcal{G}_j). \tag{6.3.4}$$

Then $Y_t(\mathcal{A}_i \cap \mathcal{G}_j)$ can be estimated by partitioning the disease count in region $\mathcal{A}_i$, $Y_t(\mathcal{A}_i)$ into the m grid square intersections, $\{Y_t(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y_t(\mathcal{A}_i \cap \mathcal{G}_m)\}$, via a multinomial sampling step as follows:

$$[Y_t(\mathcal{A}_i \cap \mathcal{G}_1), ..., Y_t(\mathcal{A}_i \cap \mathcal{G}_m)] \sim \text{Multinomial}(n = Y_t(\mathcal{A}_i) | \omega_{i1}, ..., \omega_{im}). \tag{6.3.5}$$

Then $Y_t(\mathcal{A}_i \cap \mathcal{G}_j)$ are combined via (6.3.4) to obtain the estimate of $Y_t(\mathcal{G}_j)$ for each grid square $\mathcal{G}_j$. In addition, the weights ω_{ij} which are the probability that a disease event in region $\mathcal{A}_i$ is assigned to the intersection area $(\mathcal{A}_i \cap \mathcal{G}_j)$, need to be specified. This weight should depend on two quantities, the first being the size of the intersection area between region $\mathcal{A}_i$ and grid square $\mathcal{G}_j$, $a(\mathcal{A}_i \cap \mathcal{G}_j)$, compared to the other grid squares areas of intersection $\left(\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}\right)$, which is the proportion of grid square $\mathcal{G}_j$ that lies in region $\mathcal{A}_i$. The second is the number of disease cases one would expect to occur in

grid square $\mathcal{G}_j$, that is $\mathbb{E}[Y_t(\mathcal{G}_j)] = e_t(\mathcal{G}_j)R_t(\mathcal{G}_j)$, where $R_t(\mathcal{G}_j)$ is the estimated disease risk in grid square $\mathcal{G}_j$, which is unknown. Here I estimate $R_t(\mathcal{G}_j)$ separately for each year via a purely spatial kriging approach instead of using a joint space and time kriging approach, because the later approach assumes the trends are smooth (correlated) over time. One of the goals of this chapter is to estimate the disease trends in Glasgow, therefore any additional smoothing in this initial step should be avoided. Full details of the kriging model used are given in Section 2.3.5. Then I can use the kriged estimates of disease risk in grid square $\mathcal{G}_j$, $\hat{R}_t(\mathcal{G}_j)$, to calculate the multinomial weights as follows:

$$\omega_{ij} = \frac{e_t(\mathcal{G}_j)\hat{R}_t(\mathcal{G}_j)\frac{a(\mathcal{A}_i \cap \mathcal{G}_j)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_j)}}{\sum_{k=1}^{m} e_t(\mathcal{G}_k)\hat{R}_t(\mathcal{G}_k)\frac{a(\mathcal{A}_i \cap \mathcal{G}_k)}{\sum_{q=1}^{n} a(\mathcal{A}_q \cap \mathcal{G}_k)}}, \tag{6.3.6}$$

where the denominator is included to ensure $\sum_{j=1}^{m} w_{ij} = 1$ for all $i = 1, ..., n$.

Multiple imputation algorithm

1. Generate $Y_t^{(L)}(\mathcal{G}_j)$ for $L = 1, 2, ..., l$ based on $Y_t(\mathcal{A}_i)$ via multinomial sampling steps. Here ten datasets ($l = 10$) are generated to estimate disease risk at the grid level.

2. For each imputed dataset from the previous step, fit the spatio-temporal model proposed by Bernardinelli et al. (1995) to obtain the posterior samples for all model parameters.

3. Combine the results from step 2 in order to make a model inference.

6.4 Spatio-temporal modelling at the grid level

To estimate the disease risks across the Greater Glasgow and Clyde Health Board I use the spatio-temporal model proposed by Bernardinelli et al. (1995). This model is appropriate for a few reasons. Firstly, this model assumes that the change in disease risk over time in each areal unit can be described by a linear relationship, which is suitable for the trend in these data since I have only four time points for each areal unit and therefore more complex trends are not advisable. Furthermore, the linear predictor contains separate intercepts and temporal slopes, which allows for different

risk trends in each area. In other words, the model can estimate the linear trend in disease risk for Glasgow overall, whilst also allowing for separate trends in each area, thus answering the questions of interest in Section 6.1. This model takes the form:

$$\begin{aligned} Y_t(\mathcal{G}_j)|e_t(\mathcal{G}_j), R_t(\mathcal{G}_j) &\sim \text{Poisson}[e_t(\mathcal{G}_j)R_t(\mathcal{G}_j)] \\ \ln[R_t(\mathcal{G}_j)] &= [\alpha + \phi(\mathcal{G}_j)] + [\beta + \delta(\mathcal{G}_j)] \left(\frac{t-\bar{t}}{N}\right), \end{aligned} \tag{6.4.1}$$

where $Y_t(\mathcal{G}_j)$ and $e_t(\mathcal{G}_j)$ respectively represent the observed and expected disease counts for grid square $\mathcal{G}_j$ at time t for $t = 1, ..., N$, which are estimated in the previous section. $R_t(\mathcal{G}_j)$ denotes disease risk for grid square $\mathcal{G}_j$ at time t. The global intercept term α is a fixed effect that is common for all grid squares, while β is the overall time effect (slope) and is also a fixed effect common to all grid squares. The random effects terms $\phi(\mathcal{G}_j)$ and $\delta(\mathcal{G}_j)$ represent grid square specific intercepts and slopes respectively. In other words, the intercept for grid square $\mathcal{G}_j$ can be computed by the sum $\alpha + \phi(\mathcal{G}_j)$, while the slope or trend for grid square $\mathcal{G}_j$ is the sum $\beta+\delta(\mathcal{G}_j)$. In both cases the random effect terms sum to zero to aid parameter identifiability, that is $\sum_{j=1}^{m} \phi(\mathcal{G}_j) = \sum_{j=1}^{m} \delta(\mathcal{G}_j) = 0$. Here, $\bar{t} = (1/N)\sum_{t=1}^{N} t$ is the time point average, and the term $(t - \bar{t})/N$, is used to centre time points to ensure that the intercept term represents the average disease risk over time. The random effect terms are modelled using the Leroux CAR prior (Leroux et al., 2000) given by

$$\begin{aligned} \phi(\mathcal{G}_j)|\boldsymbol{\phi}(\mathcal{G}_{-j}) &\sim \text{N}\left(\frac{\rho\sum_{k=1}^{m} w_{kj}\phi(\mathcal{G}_k)}{\rho\sum_{k=1}^{m} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{k=1}^{m} w_{kj} + 1 - \rho}\right) \\ \delta(\mathcal{G}_j)|\boldsymbol{\delta}(\mathcal{G}_{-j}) &\sim \text{N}\left(\frac{\lambda\sum_{k=1}^{m} w_{kj}\delta(\mathcal{G}_k)}{\lambda\sum_{k=1}^{m} w_{kj} + 1 - \lambda}, \frac{\sigma^2}{\lambda\sum_{k=1}^{m} w_{kj} + 1 - \lambda}\right) \\ \tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(1, 0.01) \\ \rho, \lambda &\sim \text{Uniform}(0, 1), \end{aligned} \tag{6.4.2}$$

where $\boldsymbol{\phi}(\mathcal{G}_{-j}) = [\phi(\mathcal{G}_1), ..., \phi(\mathcal{G}_{j-1}), \phi(\mathcal{G}_{j+1}), ..., \phi(\mathcal{G}_m)]$ and $\boldsymbol{\delta}(\mathcal{G}_{-j}) = [\delta(\mathcal{G}_1), ..., \delta(\mathcal{G}_{j-1}), \delta(\mathcal{G}_{j+1}), ..., \delta(\mathcal{G}_m)]$. The spatial autocorrelation parameters (ρ, λ), which control the spatial smoothness in the intercepts and slopes, are common to all time points. Spatial autocorrelation is induced via an $m \times m$ neighbourhood

matrix, $\boldsymbol{W}$. Note that I use the 4-nearest neighbours specification and set $w_{kj} = 1$ if grid square $\mathcal{G}_k$ is one of the four nearest neighbours of grid square $\mathcal{G}_j$, and $w_{kj} = 0$ otherwise. The 4-nearest neighbours specification is chosen because some grid squares have no neighbours since the grid squares with zero population have been removed, as discussed in more detail in Section 4.2.4. However, the neighbourhood matrix for the areal unit level is constructed via the most commonly used sharing common border where set $w_{ki} = 1$ if area $\mathcal{A}_k$ shares a common border with area $\mathcal{A}_i$ and $w_{ki} = 0$ otherwise.

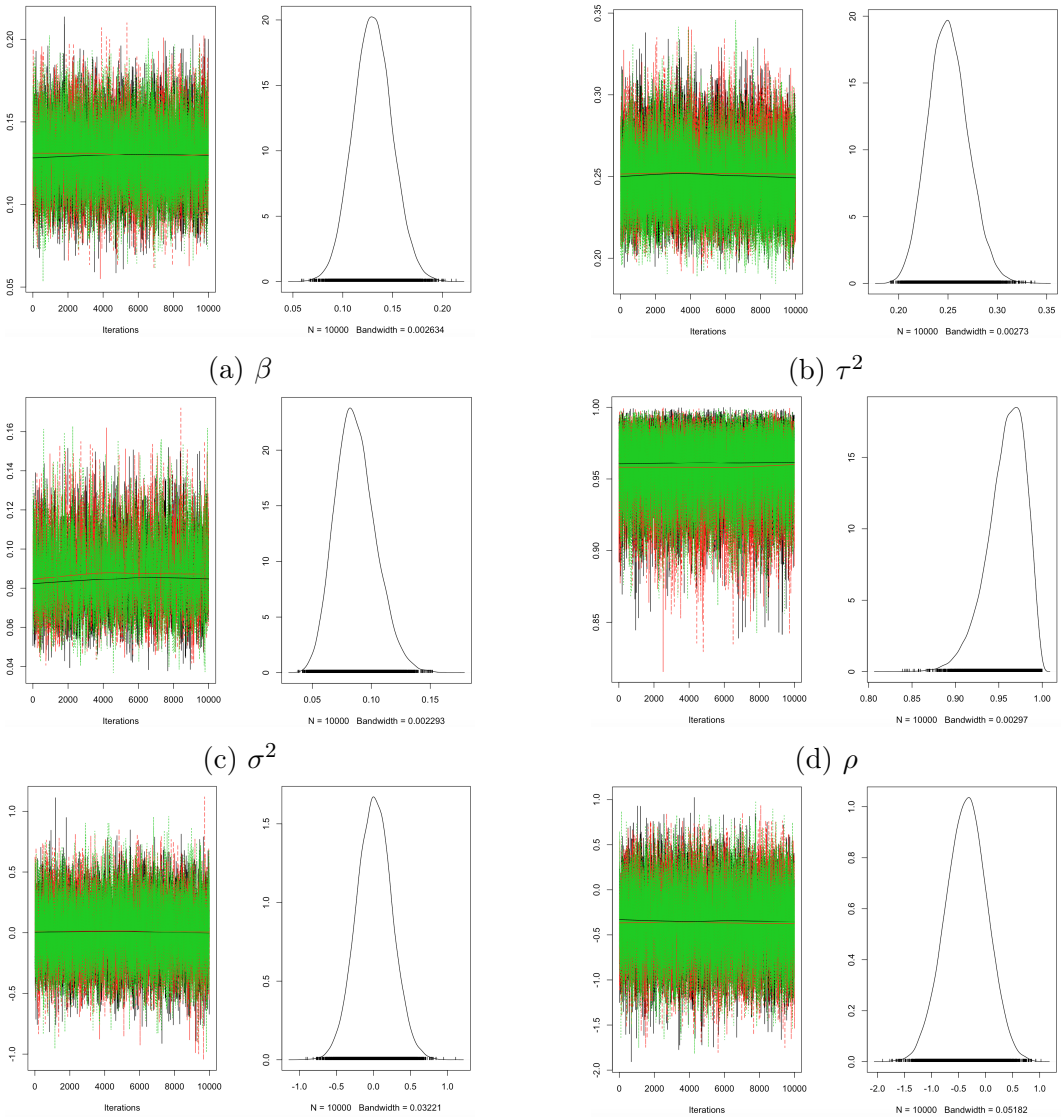
6.5 Results

The above described methodology is applied to the data outlined in Section 6.2 with the grid square sides of lengths 1,000 and 500 metres as in the previous chapters, and these results are compared against the commonly used IZ level analysis. The latter uses the same model as defined in Section 6.4, but it is applied to the areal unit level data $[Y_t(\mathcal{A}_i), e_t(\mathcal{A}_i)]$ rather than the grid level estimated data $[Y_t(\mathcal{G}_j), e_t(\mathcal{G}_j)]$. The respiratory disease risk in each area or grid square can be estimated by using the multiple imputation approach with ten imputed datasets. The Markov chain Monte Carlo is used to obtain the results, and the model runs three times to generate MCMC samples for three independent Markov chains. The model inferences are based on 200,000 iterations with a burn-in period of 50,000 and thinned by 15 for each chain. This resulting in 300,000 samples for model inference overall with 10,000 samples for each chain and 30,000 samples for each of imputed datasets. The MCMC algorithm is implemented using the CARBayesST package (Lee et al., 2018) in R (R Core Team, 2014).

6.5.1 Convergence diagnostic

The method of Gelman-Rubin (Gelman and Rubin, 1992) and trace plots assessment are used to diagnose the convergence of the posterior samples. Figures 6.5 and 6.6 illustrate trace plots of selected parameters from the models with two different grid sizes from one imputed dataset. The figures show that there is no clear pattern in the plots which indicates that the chains appear to have converged. Furthermore, the trace

plots for the other parameters and other nine datasets are very similar, therefore they are not shown. The results from the Gelman-Rubin are less than 1.1 for all selected parameters, which indicate good mixing of the chain.

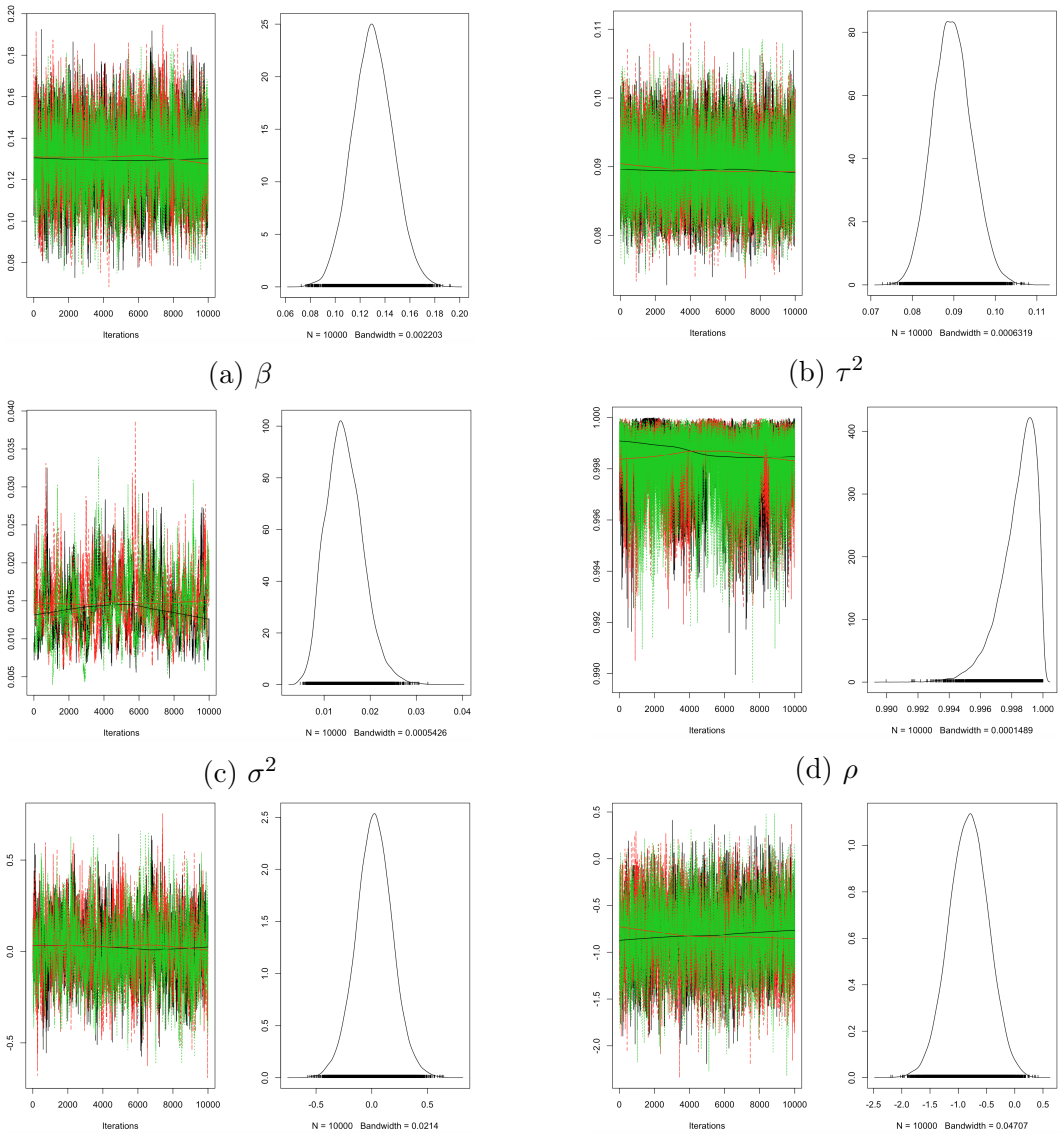


(a) β (b) τ^2

(c) σ^2 (d) ρ

(e) $\delta(\mathcal{G}_1)$ (f) $\phi(\mathcal{G}_1)$

Figure 6.5: Traceplots of MCMC samples for selected parameter (grid of size 1,000m).



(a) β (b) τ^2

(c) σ^2 (d) ρ

(e) $\delta(\mathcal{G}_1)$ (f) $\phi(\mathcal{G}_1)$

Figure 6.6: Traceplots of MCMC samples for selected parameters (grid of size 500m).

6.5.2 Sensitivity analysis

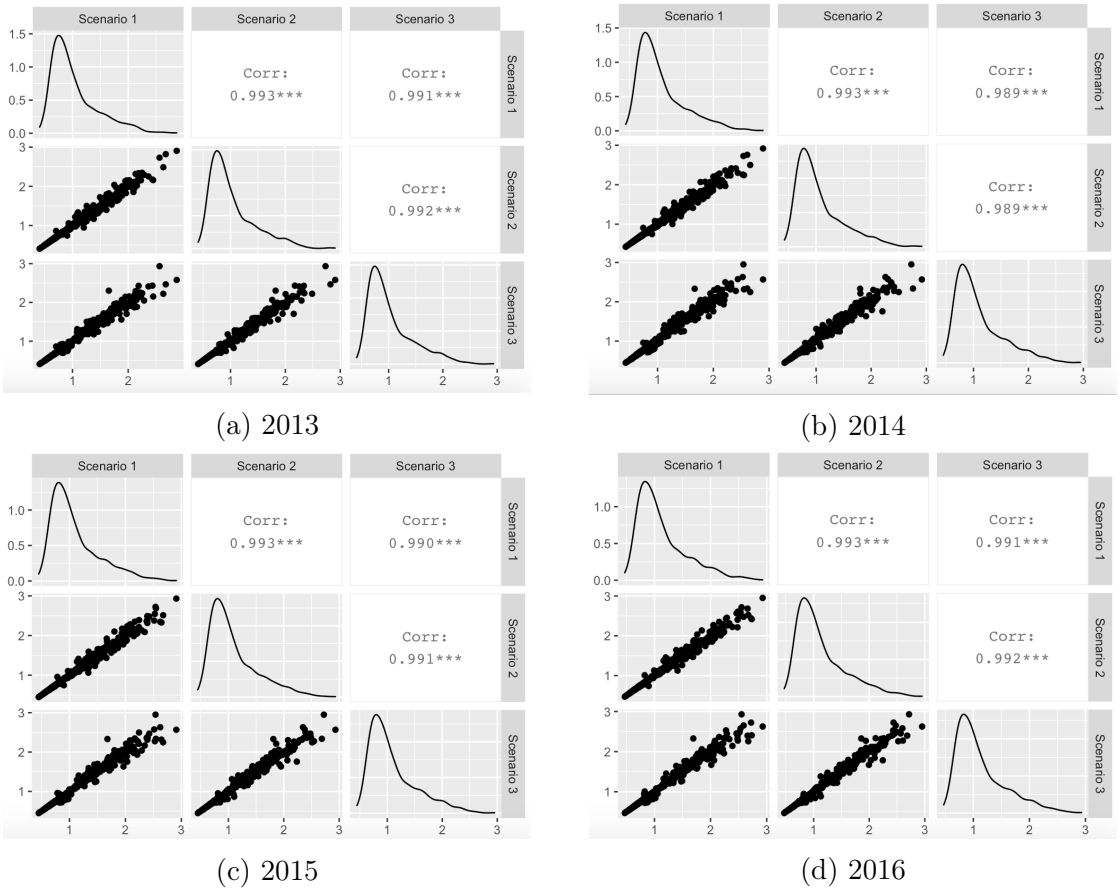
In order to examine that the posterior samples are not affected by the choices of hyperparameter, three sets of hyperpriors for the variances of random effects (τ^2, σ^2) from 6.4.2 are selected for sensitivity analysis as follows:

1. Scenario 1 - $\tau^2, \sigma^2 \sim$ Inverse-Gamma$(1, 0.01)$.

2. Scenario 2 - $\tau^2, \sigma^2 \sim$ Inverse-Gamma$(0.01, 0.01)$.

3. Scenario 3 - $\tau^2, \sigma^2 \sim$ Inverse-Gamma$(0.05, 0.0005)$.

Figures 6.7 and 6.8 show the relationship plots of the estimated risks among scenarios in the years 2013 to 2016 for grid squares of sizes 1,000 and 500 metres. The figures present that the estimated risks for each year lie on the straight line across scenarios, which indicates that different hyperpriors do not change the posterior distributions. Therefore only one hyperprior setting is used for the model inference, here Scenario 1 is randomly selected.



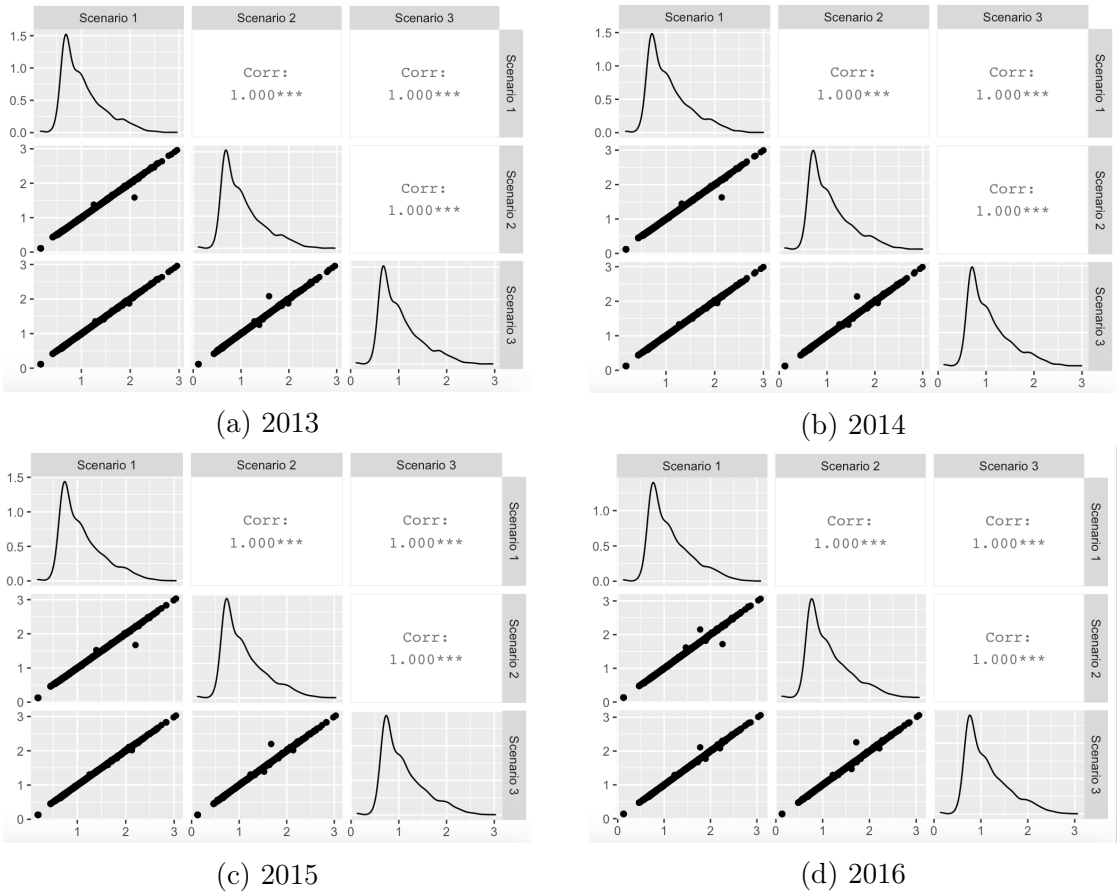
(a) 2013 (b) 2014

(c) 2015 (d) 2016

Figure 6.7: The estimated risks scatter plots of scenarios 1 - 3 for the years 2013 - 2016 (grid of size 1,000m).

6.5.3 Posterior predictive check

The predictive posterior checking is conducted in order to investigate an appropriateness of the models to the data. Since the true data at the grid level are unknown, therefore the simulated grid level data (from the model) are aggregated to the areal unit level and compare to the observed data at the areal unit level. Figure 6.9 indicates that the aggregated data and the observed data at the grid level are similar since the



(a) 2013

(b) 2014

(c) 2015

(d) 2016

Figure 6.8: The estimated risks scatter plots of scenarios 1 - 3 of Models 2 and 3 (grid of size 500m).

data fairly lie on the straight line for both grid square sizes. These results suggest that the model fits the data quite well.

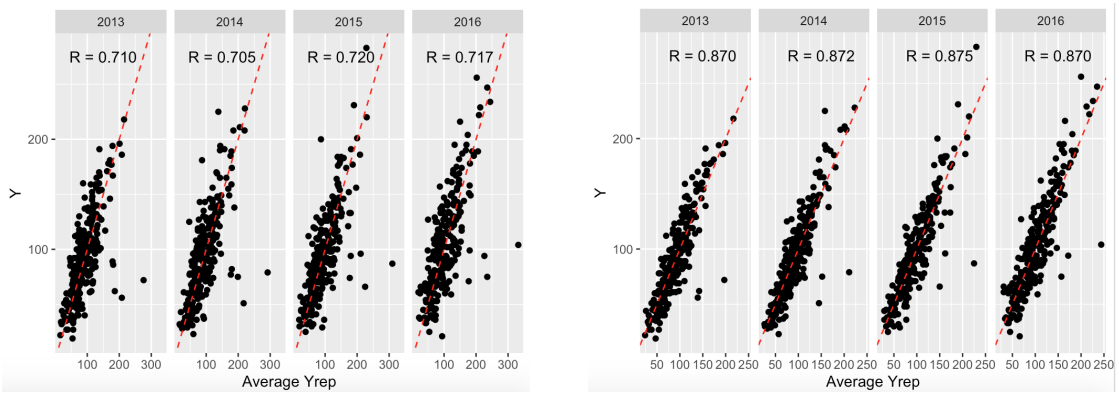
6.5.4 Main results

Recall that the three questions that I would like to address are:

i) What is the average trend over time of respiratory disease risk across the Greater Glasgow and Clyde Health Board area?

ii) How has the respiratory disease risk in each part of Glasgow changed over time in the Greater Glasgow and Clyde Health Board area from 2013 - 2016?

iii) How are the health inequalities changing over time in the Greater Glasgow and Clyde Health Board area for respiratory disease risk?



(a) grid of size 1,000m (b) grid of size 500m

Figure 6.9: Posterior predictive checks.

Figures 6.10 to 6.12 present the estimated respiratory disease risk in each area across the Greater Glasgow and Clyde Health Board between the years 2013 and 2016 at the IZ scale and grid scales (1,000 and 500 metres). Overall the spatial patterns of the estimated disease risks are similar for all scales and for each year. The correlation coefficients between the risk estimates for each year are between 0.94 and 0.99, suggesting very similar spatial surfaces each year for all three spatial scales. The areas with higher risk are amongst the most deprived areas in Glasgow, for example Clydebank and Paisley. In contrast, the areas with lower risk correspond to the affluent areas such as Clarkston and Newton Mearns.

Parameters estimation

Table 6.1 shows the estimates for the parameters from the model of Bernardinelli et al. (1995) given by (6.4.1) and (6.4.2) and their 95% credible interval at the IZ level and grid square level with grid sides of lengths 1,000 and 500 metres. The estimates are obtained from the median of the posterior samples for each parameter, and the 95% credible intervals are taken from the 2.5% and 97.5% of their posterior samples. Overall the variances of the random effects (τ^2, σ^2) at the IZ level are greater than at the grid square level (both sizes), which means that the spatial patterns in the risks and risk trends are more similar to those in neighbouring units at the grid square scale compared to the IZ level. This is because the finer spatial scales related to neighbours that are closer together, hence are more similar. The estimates for ρ are larger at the

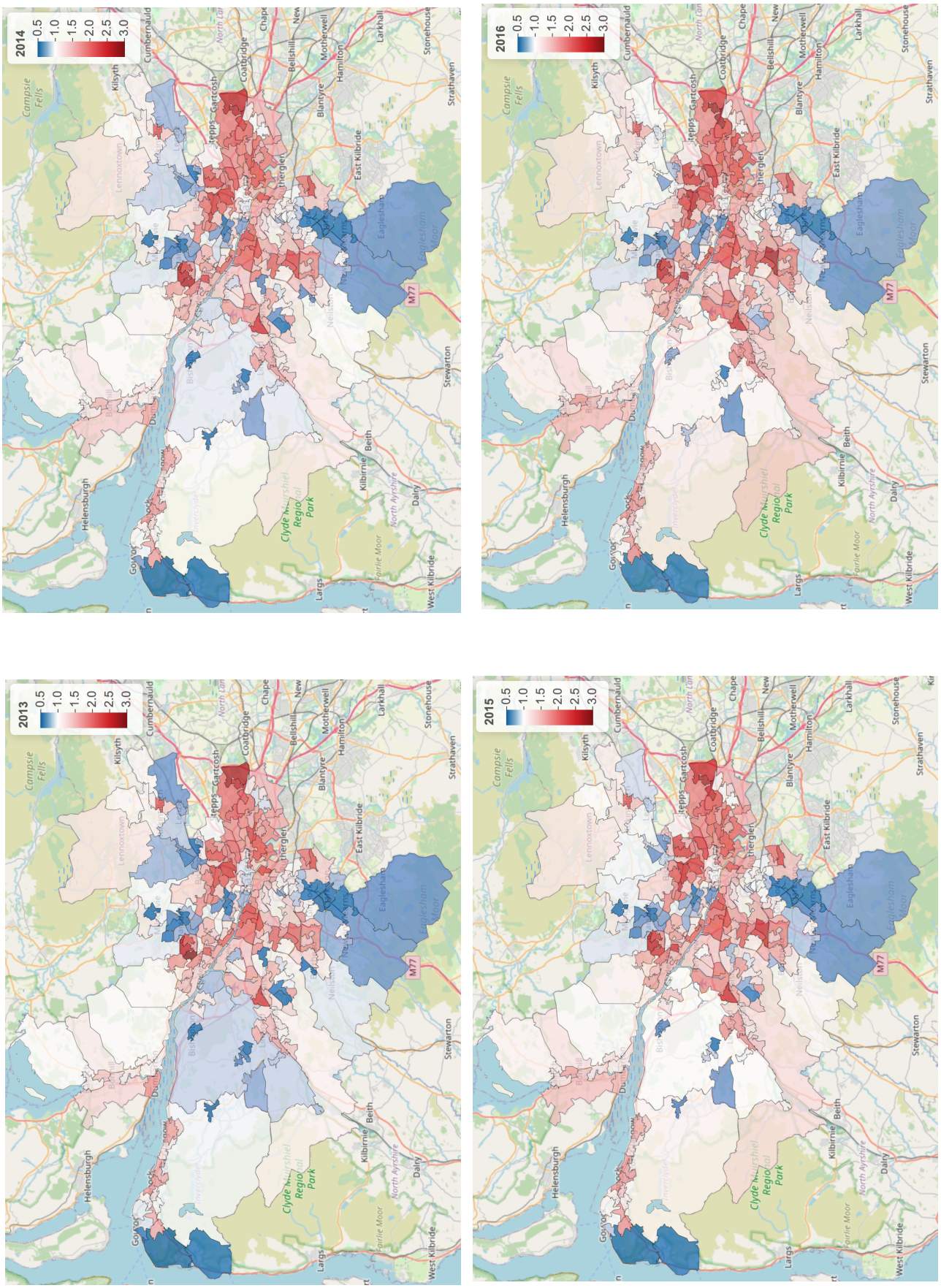


Figure 6.10: Estimated respiratory disease risk maps at the IZ level over Glasgow from 2013 - 2016.

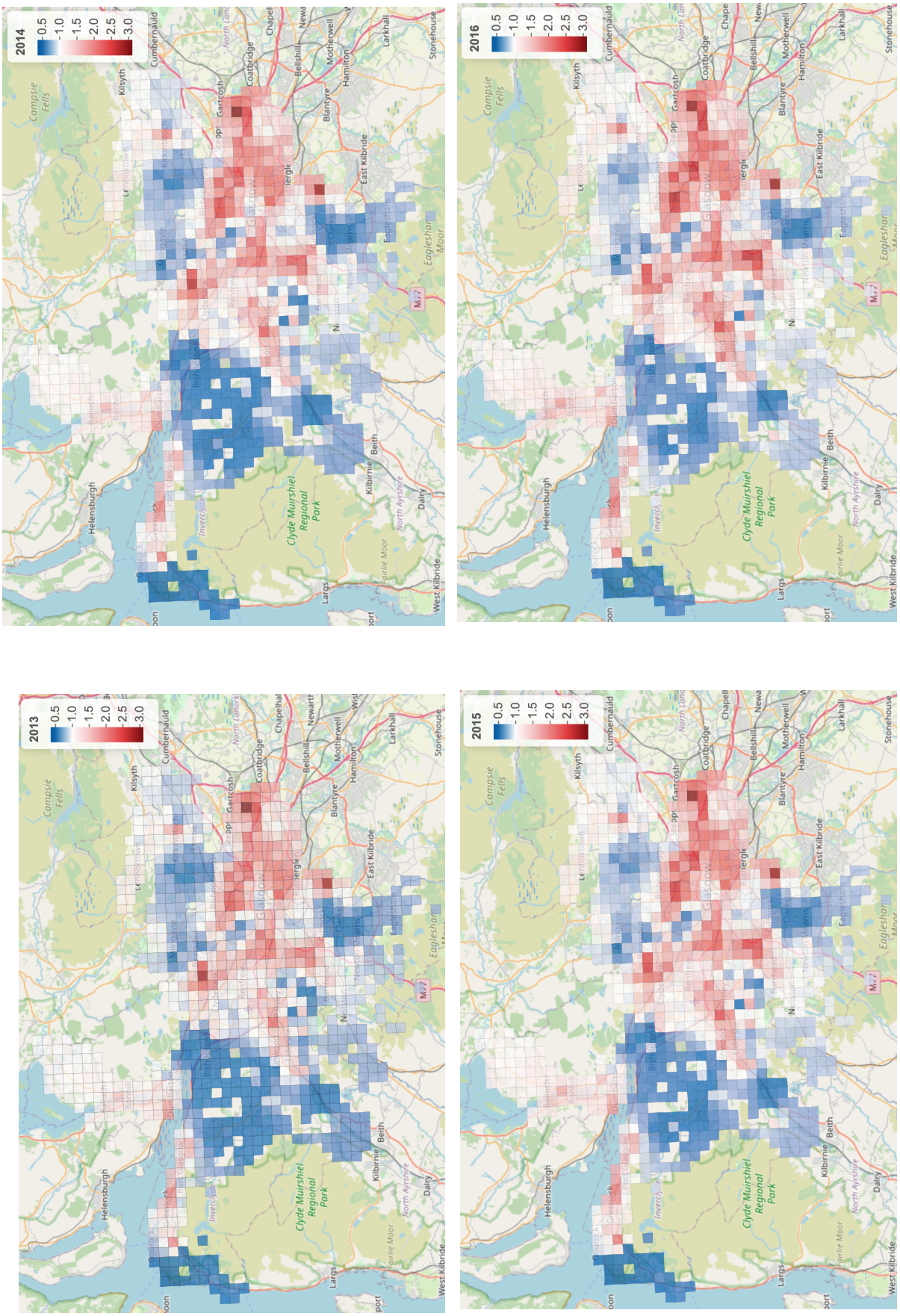


Figure 6.11: Estimated respiratory disease risk maps at the grid level (1,000 m) over Glasgow from 2013 - 2016.

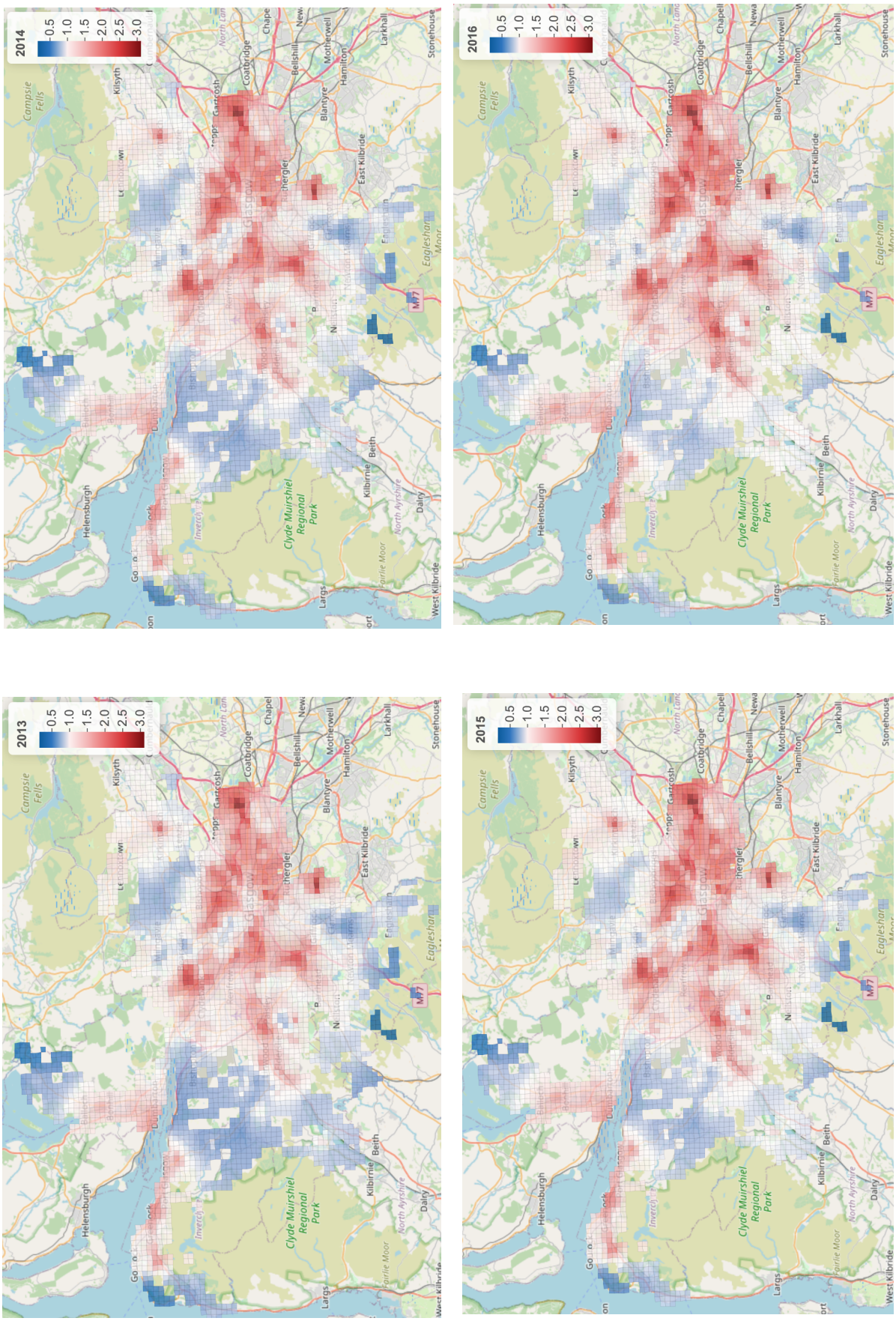


Figure 6.12: Estimated respiratory disease risk maps at the grid level (500 m) over Glasgow from 2013 - 2016.

Table 6.1: Parameters estimates and their 95% credible interval at the areal unit level and grid level (1,000 and 500 metres)

Parameter	Scale		
	IZ	**1000m**	**500m**
τ^2	0.263 (0.209, 0.327)	0.188 (0.161, 0.220)	0.090 (0.081, 0.099)
σ^2	0.069 (0.042, 0.114)	0.022 (0.012, 0.036)	0.014 (0.008, 0.022)
ρ	0.763 (0.559, 0.928)	0.979 (0.944, 0.997)	0.999 (0.995, 1.000)
λ	0.136 (0.105, 0.467)	0.949 (0.805, 0.994)	0.991 (0.961, 0.999)

grid scales than the IZ scale, suggesting on increased level of spatial smoothness due to the units being smaller and hence closer together. Furthermore, the corresponding estimates of the spatial autocorrelations (λ) in the time trends at the IZ level is only 0.136 compared to 0.991 for grid squares with sides of length 500 metres. This is again because neighbouring units at the grid level are closer together than those at the IZ level, and hence have more similar risk trends.

Temporal pattern over time

In order to answer Questions (i) and (ii) in Section 6.5, the relative risk (RR) is used to measure how respiratory disease risk has changed in the Greater Glasgow and Clyde Health Board area from 2013 to 2016. Here, I compute the relative risk for each grid square $\mathcal{G}_j$ from year t to the next year $t+1$ to estimate the relative change in risk for a one year change, $RR_{(t,t+1)}(\mathcal{G}_j)$. Based on the Bernardinelli model (6.4.1), the $RR_{(t,t+1)}(\mathcal{G}_j)$ is computed as

$$
\begin{aligned}
RR_{(t,t+1)}(\mathcal{G}_j) &= \frac{R_{(t+1)}(\mathcal{G}_j)}{R_t(\mathcal{G}_j)} \\
&= \frac{\exp[\alpha+\phi(\mathcal{G}_j)]\exp\left[(\beta+\delta(\mathcal{G}_j))\left(\frac{t+1-\bar{t}}{N}\right)\right]}{\exp[\alpha+\phi(\mathcal{G}_j)]\exp\left[(\beta+\delta(\mathcal{G}_j))\left(\frac{t-\bar{t}}{N}\right)\right]} \\
&= \frac{\exp[\alpha]\exp[\phi(\mathcal{G}_j)]\exp\left[\beta\left(\frac{t+1-\bar{t}}{N}\right)\right]\exp\left[\delta(\mathcal{G}_j)\left(\frac{t+1-\bar{t}}{N}\right)\right]}{\exp[\alpha]\exp[\phi(\mathcal{G}_j)]\exp\left[\beta\left(\frac{t-\bar{t}}{N}\right)\right]\exp\left[\delta(\mathcal{G}_j)\left(\frac{t-\bar{t}}{N}\right)\right]} \\
&= \frac{\exp\left[\frac{\beta t}{N}\right]\exp\left[\frac{\beta}{N}\right]\exp\left[-\frac{\beta\bar{t}}{N}\right]\exp\left[\delta(\mathcal{G}_j)\left(\frac{t}{N}\right)\right]\exp\left[\frac{\delta(\mathcal{G}_j)}{N}\right]\exp\left[\delta(\mathcal{G}_j)\left(-\frac{\bar{t}}{N}\right)\right]}{\exp\left[\frac{\beta t}{N}\right]\exp\left[-\frac{\beta\bar{t}}{N}\right]\exp\left[\delta(\mathcal{G}_j)\left(\frac{t}{N}\right)\right]\exp\left[\delta(\mathcal{G}_j)\left(-\frac{\bar{t}}{N}\right)\right]} \\
&= \exp\left[\frac{\beta}{N}\right]\exp\left[\frac{\delta(\mathcal{G}_j)}{N}\right].
\end{aligned}
\tag{6.5.1}
$$

This relative risk indicates the yearly rate change in the risk for respiratory disease in each grid square. Therefore the overall trend in respiratory disease risk over time [Question (i)] can be computed by averaging the relative risks over all grid squares via $\frac{1}{m}\sum_{i=1}^{m}\left(\exp\left[\frac{\beta}{N}\right]\exp\left[\frac{\delta(\mathcal{G}_j)}{N}\right]\right)$. The posterior distribution of this quantity can be obtained by firstly computing the relative risk values from (6.5.1) for all grid squares and posterior samples. Then compute the mean of those values over all grid squares as above for each MCMC sample. Then compute the posterior median as a point estimate and take the 2.5 and 97.5 percentiles as the 95% credible interval. Here I compare the relative risk at the three scales which are IZ scale and grid scales of sides 1,000 and 500 metres. For the relative risk at the IZ level, it can be computed in the same manner as the grid level via (6.5.1) but using the IZ data instead of grid data. Overall they produce similar results which are presented in Table 6.2. The point estimate of the yearly rate change in respiratory disease risk across the Greater Glasgow and Clyde Health Board at the IZ level is 1.029 which means that the respiratory disease risk in Glasgow is increasing by 2.9% every year. While the rate change at the grid level for both sizes are similar which is approximately 1.04. In other words, the risk is rising by 4% which is slightly higher than the IZ level. However these increasing between two scales are not statistically significant since their 95% CI overlap.

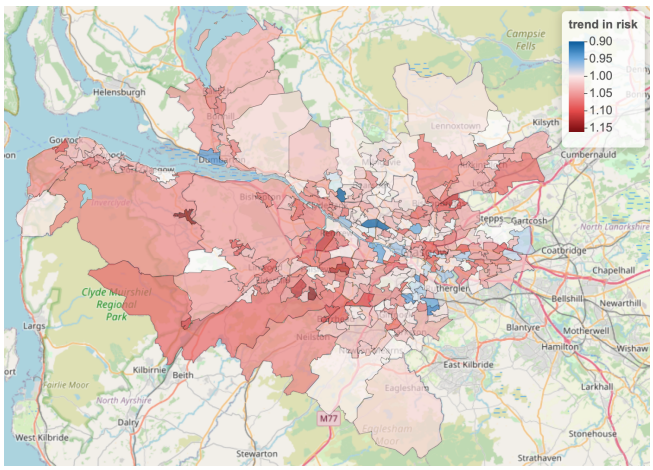
Table 6.2: Annual changes in disease rates and their 95% credible interval.

Scale	Relative risk	
IZ	1.029	(1.023, 1.036)
1,000 m	1.042	(1.034, 1.051)
500 m	1.043	(1.035, 1.052)

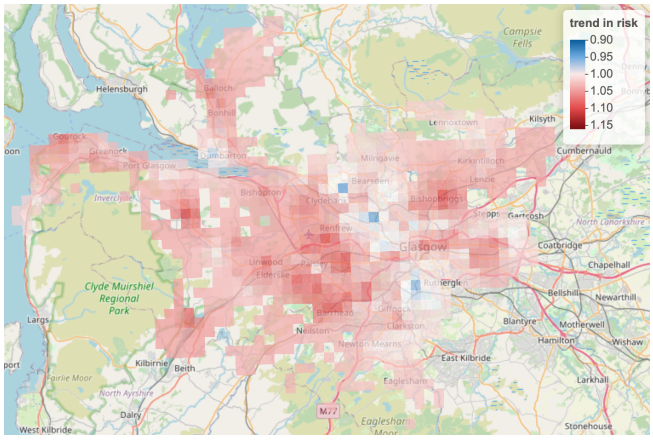
In addition, the annual rate changes in the risk of respiratory disease in each area [Question (ii)] are illustrated in Figure 6.13 which includes all three scales; IZ level and grid square sides of lengths 1,000 and 500 metres. The spatial maps show similar patterns of respiratory disease risk trend in Glasgow for all three scales. However, the IZ level produces a discrete spatial map since it is assumed that the risk is constant within each IZ. While the grid sides of lengths 1,000 and 500 metres produce closer approximation to a pseudo continuous map over Glasgow with the latter being a closer approximation. The spatial maps show that there are areas with increasing ($RR_{(t,t+1)} > 1$) respiratory disease risk trends and also areas with decreasing ($RR_{(t,t+1)} < 1$) trends. The IZ level produces the biggest range in the risk trend (0.900 - 1.153), while the grid scales at 1,000 and 500 metres produce smaller ranges by (0.948 - 1.111) and (0.983 - 1.071) respectively. Furthermore 82.88% of the IZs have an increased trend in the risk, while 98.36% for the grid square size 1,000 metres and 98.97% for the grid square size 500 metres exhibit increased trends. However, these increases are considered statistical significance by 22.57% for the IZ level and 28.02% and 32.03% for the grid squares of sizes 1,000 and 500 metres respectively. While there is no statistically significant decrease for areas with decreased trend at all scales. These results indicate that roughly 30% of all areas in Glasgow are getting worse in respiratory disease risk. In summary, the areas with highest increased trend are Paisley, Bishopbriggs and Barrhead.

Health inequalities

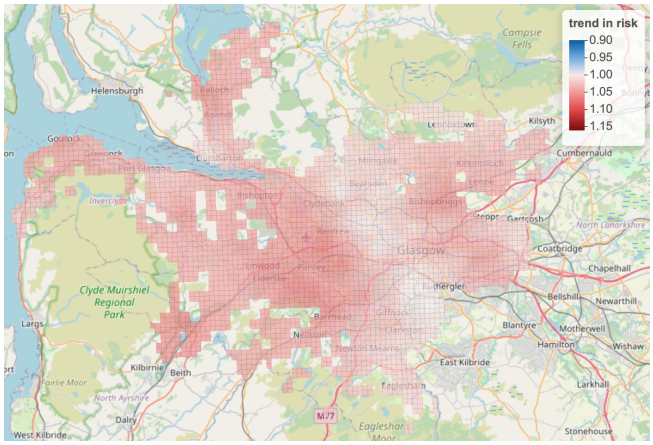
In order to examine how health inequalities change over time in the Greater Glasgow and Clyde Health Board area for respiratory disease risk [Question (iii)], Figure 6.14 shows boxplots of the yearly estimated respiratory disease risks for all three scales from 2013 to 2016. The numbers in red on the top of this figure are the spatial interquartile



(a) Intermediate zones



(b) Grid sides of length 1000 metres



(c) Grid sides of length 500 metres

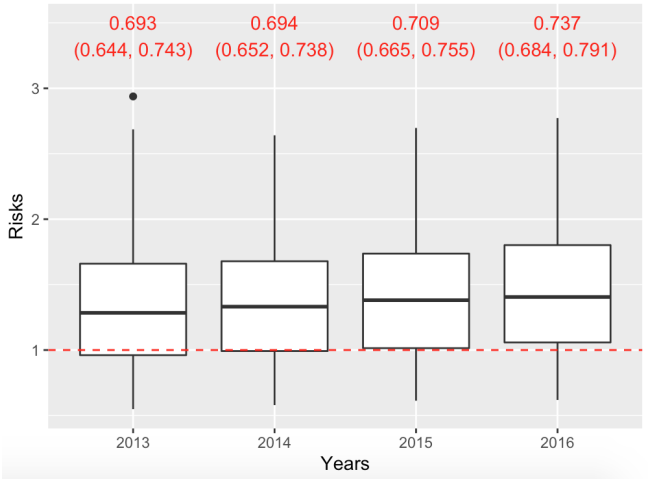
Figure 6.13: Maps of yearly rate change for respiratory disease risk across the Greater Glasgow and Clyde Health Board.

range (IQR) and their 95% credible intervals, which are used to measure the variability in the risk for each year and can be calculated as the difference between upper and lower quartiles. Therefore, I use IQR to measure health inequalities in this study. These IQR values increase over time for all scales, for example the IQR for the grid size 1,000 metres in 2013 is 0.585 compared to 0.638 for 2016. This means that health inequalities in Glasgow are increasing over time, which is a consistent finding regardless of the spatial scale considered. However, these increases are not statically significant for all scales since their 95% CIs are overlapping.

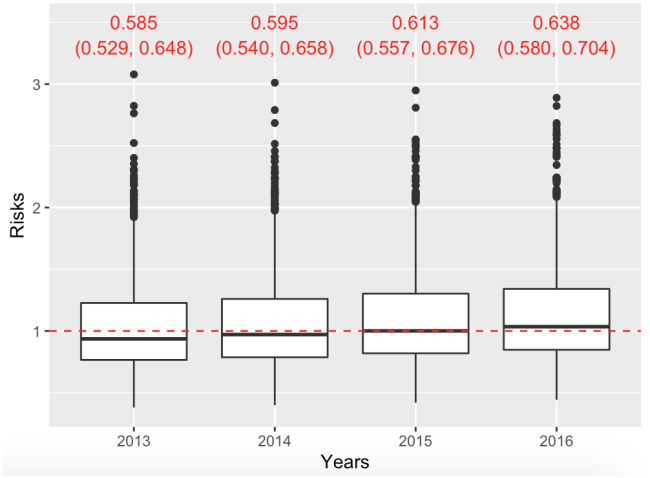
The average estimated respiratory disease risks at the grid level in each year seem to be lower than those at the IZ level as presented in Figure 6.14. This is because the areal units with low values of SIR are more likely to be geographically large rural IZs, while areas with high values of SIR tend to be geographically small urban IZs. Consequently, when the areal disease data are imputed onto the grid square level, the areas with low risk which are geographically large, have more grid squares to allocate the areal disease counts to. Meanwhile, areas with high risk often tend to be geographically small, and thus have fewer grid squares to allocate the areal disease counts to. Therefore, the proportion of grid squares which are estimated to have low risk will be much larger than the proportion of IZs which have low SIRs. This results in the median disease risk being lower when the areal unit level data are transformed to the grid level.

6.6 Conclusion

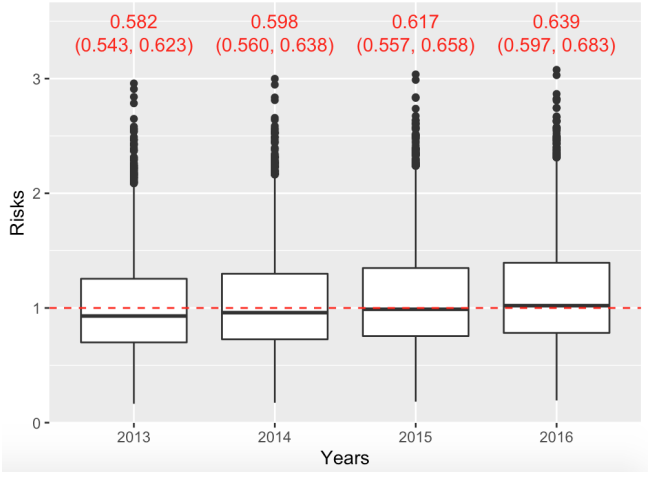
This chapter aimed to investigate the variation in respiratory disease risk at the grid level over the Greater Glasgow and Clyde Health Board, and compare the results from the grid level data with the commonly used technique to estimate disease risk at the IZ level. The spatio-temporal model used to achieve these aims is proposed by Bernardinelli et al. (1995). This model includes two pairs of parameters; one which controls the spatial pattern and one which controls the temporal trend. Each pair consists of a global fixed effect and a set of random effects which follow a conditional autoregressive model. The model was applied to data on the annual numbers of hospital admissions



(a) Intermediate zones



(b) Grid sides of length 1000 metres



(c) Grid sides of length 500 metres

Figure 6.14: Boxplots of respiratory disease risk at the grid level across the Greater Glasgow and Clyde Health Board from 2013-2016.

for respiratory disease in the Greater Glasgow and Clyde Health Board from 2013 to 2016. The results obtained at the IZ level indicate that the main drawback is the risk estimates are assumed to be constant within each IZ, which is not necessary realistic. Therefore I overcome these problems by proposing novel models which estimate the disease risk on a pseudo-continuous spatial surface using a grid based approach. This allows risk to vary within each IZ. As these grid squares get smaller this results in a pseudo-continuous spatial risk surface.

In order to make inference at the grid square level, the IZ data were transformed onto the grid level before fitting the model. The method of grid level transformation was presented in Section 6.2. Overall the estimated spatial patterns in respiratory disease risk are similar for all scales and also for each year, and the areas with higher risks tend to be the more deprived areas. This finding is similar to those obtained in Chapters 3 - 5. However the IZ level produces the discrete spatial map of the risk estimates, while the grid level produces the pseudo-continuous map which is smoother than the IZ level.

A relative risk is used to identify the overall trend in respiratory disease risk, as well as individual trends for each grid square. The overall risk for the whole region is increasing by 4% every year. Furthermore, more than 95% of all grid squares have increasing disease risks, which means that the majority of areas in Glasgow are getting worse in respiratory disease risk. Note that there are approximately 30% of grid squares are statistically significant increase over time. In order to investigate how health inequalities change over time, the IQR is computed to measure these inequalities. The results showed that health inequality in the Greater Glasgow and Clyde Health Board remains unchanged over years 2013 to 2016 due to the 95% credible intervals of IQRs are overlapping.

This methodology allows me to estimate the overall trend in disease risk, as well as individual trends for each area. However there are only four time points for each grid square, which is clearly a drawback of this study. Therefore the only relationship that could sensibly be specified to describe change in the data over time is a linear relationship, as more complex trends are not reasonable given the short temporal duration.

Future work could include more time points to obtain more accurate results. This study could also be adopted for other diseases, as a single disease does not comprehensively measure health inequalities. Therefore multiple diseases could also be considered for future study.

Chapter 7

Conclusion

The main focus of this thesis was to measure health inequalities in the Greater Glasgow and Clyde Health Board and examine how they changed over time. These issues have been around for a long time and the Scottish Government and health authorities have tried to minimise the health inequality gap. Glasgow was chosen since it has the lowest life expectancy of all major cities in the UK (Walsh et al., 2016). There have been several studies focusing on health inequality in Glasgow which were discussed in Chapter 1. Generally, the study region is split into non-overlapping areal units and then disease risk is estimated for people living in each areal unit, via different approaches e.g. SIR and CAR models. The SIR approach generally produces an unstable estimates disease risk when the population in the study areas is small or the disease being studied is rare. To tackle this problem, a Poisson generalised linear model with random effects is commonly used to estimate the disease risk in each area. These random effects account for the spatial autocorrelation that typically present in spatial data and are normally modelled via CAR models. Such an approach, however assumes a constant disease risk within each area, which is not necessarily realistic especially in the rural areas which normally have larger areal units than urban areas. In addition, boundaries of the study region are arbitrary and can be changed over time, therefore the results cannot be directly compared between these two (or multiple) sets of boundary. Moreover, this existing approach estimates the risk where no people live e.g. mountain and field, which is not reasonable. Therefore, this study aimed to overcome this challenge by creating a set of grid squares over the study region and trying to make an inference only on the grid square with non-zero population. This novel approach also overcomes the

modifiable areal unit problem which is the most common issues related to aggregated spatial data. Furthermore, when the grid squares become smaller the inference will get closer to an individual level which avoid an ecological fallacy problem.

7.1 Summary

In Chapter 3, a spatial model was proposed for quantifying health inequalities in the Greater Glasgow and Clyde Health Board at the areal unit level utilising Poisson log-linear CAR models. This chapter presented the standard Leroux CAR method that is commonly used in existing studies. This model was applied to the hospital admission numbers for respiratory disease in the years 2015 - 2016. It was found that the areas with higher risks were deprived areas such as Clydebank and Paisley. Conversely, the areas with lower risk are wealthier areas such as, Bearsden, Clarkston, and Eaglesham. These results indicated that people living in more deprived areas are more likely to be hospitalised for respiratory disease than those living in wealthier areas. This is because people in poor areas are more likely to drink, smoke, do less physical activity, and consume unhealthy food which are the main factors related to respiratory disease (Pampel et al., 2010; World Health Organization and others, 2007). There were however two main concerns in this study, which are that disease risks were estimated in areas where no people live, and a constant disease risk within each area or IZ is assumed.

Chapter 4 tackled these challenges by estimating pseudo-continuous risk patterns over the Greater Glasgow and Clyde Health Board. This approach created grid squares across the study region, and aimed to estimate disease risks for each grid square. It was not sensible to estimate the risk in areas where nobody lives, therefore grid squares with zero population were removed. The areal unit data were transformed to the grid level scale and then used to fit the spatial CAR model proposed by Leroux et al. (2000) to quantify the spatial variation in disease risk at the grid level. The expected disease counts at the grid level were estimated based on population density in each grid square, while the disease counts were estimated via multinomial sampling steps. The probabilities in the multinomial sampling depended on the expected disease counts in the intersection areas between each IZ and each grid square, which were unknown. There-

fore, they were estimated using the area of intersection between each grid square and each IZ, and the expected counts and disease risks in each grid square. The latter were unknown but were estimated by two possible methods; the naive method assumed a constant risk over all grid squares (Model 2), and the second method estimated the risks via kriging (Model 3). These proposed models were tested on their performance via a simulation study including the reference model (Model 1) which was fitted to the true simulated grid level data. Simulated data of grid squares sides of lengths 1,000 and 500 metres were used in the simulation study. According to the simulation results, Model 3 was the better model to estimate disease risk at the grid level since it produced unbiased estimates and smaller RMSE compared to Model 2. This is because the initial estimates of disease risk for each grid square in the multinomial sampling steps were estimated via kriging, which is more realistic than assuming constant risk across the study region (Model 2). Hence, the number of disease cases in each grid square were closer to the true grid level disease cases than the latter. Model 2 however produced slightly less RMSE values than Model 3 when estimating regression parameter related to unknown grid level covariate. In addition, both models produced bias estimates for regression parameter related to the true grid level covariate, therefore multiple imputation approach is not recommended when the aim of a study involves to the true grid level data. Overall, Model 3 is the better choice when multiple imputation has been chosen and it is therefore used as a comparative model in Chapter 5.

Chapter 5 introduced a novel data augmentation method to estimate the disease counts at the grid level, which allowed for uncertainty when estimating disease risk and model parameters. Data augmentation basically has two iterative steps. The first step estimates disease counts at the grid level via multinomial sampling based on aggregated disease counts at the areal unit level and the current values of the grid level model parameters. Then in the second step, all model parameters are updated based on the estimated grid level disease counts from the previous step. These two steps are iterated within an MCMC algorithm to estimate the disease risks at the grid level. There were two new models proposed in this chapter, the standard intrinsic CAR model (Model 4) was fitted to the estimated grid level data, while Model 5 estimates the variance of random effects via empirical Bayes and fixes it throughout the process in order to

reduce the variation in the estimated risks. The performance of each model was tested via a simulation study in the same manner as Chapter 4 where 100 datasets were generated in this study. Since some simulated datasets for Model 4 produced non-converged MCMC chains, leading to unstable results therefore more than 100 datasets were generated in order to obtain 100 converged datasets for this model. The results indicated that Model 4 is not recommended when data augmentation is used since it produced unstable disease risk estimates. Overall Model 5 (data augmentation) performed better than Model 3 (multiple imputation) when estimating regression parameters since it produced unbiased estimates and smaller RMSE values. While Model 3 performed slightly better when estimating disease risk at the grid level since it produced smaller RMSE values.

In Chapter 6, the spatial model was extended to measure health inequalities in the Greater Glasgow and Clyde Health Board over time via a multiple imputation approach since the results in Chapter 5 suggested that such approach is better than data augmentation if the main aim is to estimate disease risks. The spatio-temporal model used to achieve the goal was proposed by Bernardinelli et al. (1995), which is appropriate for spatial data over short time periods since the model assumes a linear trend in the disease risks over time which is all that can sensibly be estimated with so few data points over time. Therefore the model was suitable for data used in this chapter, which were yearly counts of the hospital admission number for respiratory disease between the years 2013 and 2016 (4 time points).

7.2 Main findings

The novel approach introduced in this thesis was pseudo-continuous grid level disease risk modelling using both multiple imputation and data augmentation approaches. Each approach has advantages in different circumstances according to the simulation results. Data augmentation performed better when estimating regression parameters in both types of covariate which are the true known covariate at the grid level and a covariate that has been disaggregated to the grid level. While multiple imputation performed better when estimating disease risks at the grid level. Therefore the choice

of selection between these two approaches should be made based on the objectives of the study. Data augmentation is likely to perform better when the main aim of the study is to identify factors related to the disease risks in each area, while the multiple imputation approach is likely to be more appropriate if the estimation of disease risk is the main purpose.

Grid squares size of 500 metres generally produced similar risk patterns to the grid squares size of 1,000 metres, however the latter produced better estimates for all model parameters and disease risks at the grid level regarding RMSE values across scenarios. This is because the grid square of size 1,000 metres have fewer data points to be estimated than grid squares of size 500 metres (853 vs 3,106 grid squares), therefore the results were more accurate than the latter. However, the finer grid squares produced closer to continuous disease maps, which means that when the grid size enlarges it causes the disease maps to be more discrete and pixelated. There is a trade off between a continuous disease map and more accurate estimates of disease risk, if the grid squares are too big, they might cover two areas with completely different disease risks. Conversely, if the grid squares are too small, they might cover too many areas with homogeneous disease risks, which could lead to less accurate estimates and higher computational demand.

Each model was applied to respiratory disease data in the Greater Glasgow and Clyde Health Board area in order to illustrate the spatial variation in disease risks and identify areas of higher and lower risks. Both approaches produced similar patterns of estimated disease risks. The areas of higher risks are in the east and the north of Glasgow city centre such as Easterhouse, Drumchapel and Possilpark, which are deprived areas of the city. In contrast, the areas with lower risks are in south-west and west of the city centre such as Newton Mearns, Kelvinside and Jordanhill which are prosperous parts of the city. These results suggested that people living in poorer areas are facing higher risks than those living in more affluent areas. Furthermore, the results from the spatio-temporal model in Chapter 6 indicated that the overall disease risk was increased by 4% per year, and approximately 30% of the areas in Glasgow were statistically significant increased in the disease trend which means that they were getting worse in terms

of respiratory disease. This is corresponding to the trend in respiratory mortality rates for the years 2010 to 2016, which is slightly increasing, while other diseases (e.g. cancer, Coronary heart disease, and Cerebrovascular disease) are decreasing over time (National Records of Scotland, 2016a). It is the fact that people will die from some diseases even after health care policies have been improved. Therefore when the trends of other diseases risks decrease, it is more likely that respiratory disease risk increases over time. Moreover, health inequality in respiratory disease in Glasgow was also slightly widening every year which is measured by IQR. However these inequality is not statistically significant increases due to their 95% credible interval are overlapping. This result corresponds to the Scottish health survey in 2016 (National Records of Scotland, 2016b) which shows that the gaps in the amount of tobacco smoking and alcohol consumption between people living in the most and least deprived areas are widening in Scotland.

7.3 Limitations and future work

There are some limitations to this study, both in terms of the proposed methods and our data. From a data perspective, there were only four time points (2013 - 2016) used in Chapter 6 which were purchased from NHS Scotland. Thus, the only trend in disease risks over time that can be assumed was a linear relationship. If the data could be obtained over a longer period of time, a more complex relationship could be considered as well as a choice of spatio-temporal models. Furthermore, our disease data are in the same original IZs every year but these novel methods can be utilised to solve the modifiable unit areal unit problem that was described in Chapter 4, where the boundaries of the IZs can be changed over time. Specifically, in the year 2011 the Scottish Government decided to redraw the boundaries of the data zones and therefore if the disease data could be obtained before 2013, our novel approaches could also be used to carry out comparable inference across both set of IZs, something which has not previously been easily possible.

Future work could involve extending these methods to larger areas e.g. Scotland and the UK. However, there were already 3,106 grid squares (size of 500 metres) in the

Greater Glasgow and Clyde Health Board alone and therefore the larger area scale could lead to a huge number of grid squares. This could lead to computational infeasibility since there would be many data points and model parameters being estimated. This is one reason that only one health board area was selected in this study. However one way that we could overcome this problem is adopting an approach with non-uniform grid square sizes. Specifically, IZs with small areas (e.g. in the cities) could be given smaller grid squares (e.g. 500 metres), while IZs with large areas (e.g. rural) could be given larger grid squares (e.g. 5,000 metres). This also reduces a problem relating to computational demand. Note that the grids would not necessarily be a regular grid square (e.g. rectangle).

For the multiple imputation approach, disease counts in each grid square were estimated once before fitting a spatial model as described in Chapter 4. Therefore this approach does not allow for uncertainty in disease counts at the grid square level, hence if the estimated grid level disease counts are not accurate, this could lead to unreliable results. However, this issue was tackled by the data augmentation approach which is outlined in Chapter 5 by updating the disease cases and model parameters in the MCMC steps simultaneously. These approaches can be applied to disease counts relating to non-overlapping areas (e.g. IZ) since the estimation of grid level data are based on the intersection areas between each grid square and each IZ. However, they could not be utilised for data with overlapping areas or unknown boundary areas, for example, general practice (GP) surgeries where doctors prescribe general medication and provide a prescriptions for non-severe patients. Patients do not necessarily attend the nearest surgeries, especially in urban areas where there are multiple competing surgeries in close proximity. This could be studied in more details in a future study.

A validation analysis is an essential method to confirm the acceptable model inference, however this analysis is not carried out since I do not have a second source of the data that can measure the same thing as the data used in this study. Furthermore, in this study a single disease was applied to measure health inequality in the Greater Glasgow and Clyde Health Board. However, measuring health inequalities between populations in difference areas is an extremely complex issue, therefore a single disease may not be

sufficient to identify overall health inequalities. Future work could extend this simple approach, univariate disease risk modelling to a more realistic multivariate approach to further investigate how health inequalities change over time.

In conclusion, this thesis aims to measure health inequalities in the Greater Glasgow and Clyde Health Board and investigate how they changed over time. The models used in this study are based on the conditional autoregressive models and the proposed approaches to estimate disease risks at the grid level which are multiple imputation and data augmentation. These approaches are compared via simulation studies and then applied to the real data which are the number of hospital admissions for respiratory disease. The novelty of this work is to create a pseudo - continuous risk surface across the study region in order to overcome the problems of the existing works such as the areas with zero populations being removed before estimating disease risks and constant risk in each area (intermediate zone) is not assumed. Multiple imputation outperforms other options considered when the objective of the study is to estimate disease risks at the grid level, while data augmentation is preferable for the study involving with the grid level covariate. There are however some limitations of this work which are mentioned above, as well as the future works that could overcome the limitations.

Bibliography

D. Acheson. Inequalities in health: Report on inequalities in health did give priority for steps to be tackled. *BMJ: British Medical Journal*, 317(7173), 1998.

P. D. Allison. Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, 28(3):301–309, 2000.

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

M. Aregay, A. B. Lawson, C. Faes, and R. S. Kirby. Bayesian multi-scale modeling for aggregated disease mapping data. *Statistical Methods in Medical Research*, 26(6): 2726–2742, 2017.

M. Bartley. *Health Inequality: An introduction to concepts, theories and methods Second Edition*. 11 2016. ISBN 978-0-7456-9109-1.

T. Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

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

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

D. Black, J. Morris, and P. Townstand. Inequalities in health: The Black report and the health divide. 1982.

J. Carpenter and M. Kenward. MAR methods for quantitative data. *Missing Data in Randomised Controlled Trials—a Practical Guide. Birmingham: National Institute for Health Research*, 2008.

G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985. ISSN 00031305. URL http://www.jstor.org/stable/2682801.

S. J. Dark and D. Bram. The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, 31(5):471–479, 2007.

P. J. Diggle, P. Moraga, B. Rowlingson, B. M. Taylor, et al. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

A. Ellis and R. Fry. Regional health inequalities in England. *Regional Trends*, 42(1): 60–79, 2010.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

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

J. Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991.

P. Hanlon, D. Walsh, and B. Whyte. *Let Glasgow Flourish*. Glasgow Centre for Population Health, 2006.

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

E. Jack. *Estimating the changes in health inequalities across Scotland over time*. PhD thesis, University of Glasgow, 2019.

E. Jack, D. Lee, and N. Dean. Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3):1061–1080, 2019.

H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceeding of the Royal Society of London*, pages 453–461, 1946.

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

D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

J. Law. Exploring the specifications of spatial adjacencies and weights in Bayesian spatial modeling with intrinsic conditional autoregressive priors in a small-area study of fall injuries. *AIMS Public Health*, 3(1):65, 2016.

A. Lawson and D. Lee. Bayesian disease mapping for public health. In *Handbook of Statistics*, volume 36, pages 443–481. Elsevier, 2017.

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

D. Lee. CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013.

D. Lee. A tutorial on spatio-temporal disease risk modelling in R using Markov chain Monte Carlo simulation and the CARBayesST package. *Spatial and Spatio-temporal Epidemiology*, page 100353, 2020.

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

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

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

B. Leroux, X. Lei, N. Breslow, M. Halloran, and B. D. Elizabeth. Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 116:179–191, 2000.

K. A. Levin and A. H. Leyland. A comparison of health inequalities in urban and rural Scotland. *Social Science & Medicine*, 62(6):1457–1464, 2006.

Y. Li, P. Brown, D. C. Gesink, and H. Rue. Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21(5): 479–507, 2012a.

Y. Li, P. Brown, H. Rue, M. al Maini, and P. Fortin. Spatial modelling of lupus incidence over 40 years with changes in census areas. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 61(1):99–115, 2012b. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/41430951.

W. A. Link and M. J. Eaton. On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115, 2012.

J. P. Mackenbach, I. Stirbu, A.-J. R. Roskam, M. M. Schaap, G. Menvielle, M. Leinsalu, and A. E. Kunst. Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine*, 358(23):2468–2481, 2008.

Y. C. MacNab, P. J. Farrell, P. Gustafson, and S. Wen. Estimation in Bayesian disease mapping. *Biometrics*, 60(4):865–873, 2004.

M. Marmot, J. Allen, P. Goldblatt, T. Boyce, D. McNeish, M. Grady, and I. Geddes. The Marmot review: Fair society, healthy lives. *The Strategic Review of Health Inequalities in England Post-2010*, 2010.

G. McCartney. Illustrating health inequalities in Glasgow. *Journal of Epidemiology & Community Health*, pages jech–2010, 2010.

G. McCartney. What would be sufficient to reduce health inequalities in Scotland. *Submission to the Scottish Government's Health Inequalities Taskforce. Edinburgh, NHS Health Scotland*, 2012.

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

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

C. J. Murray, E. E. Gakidou, and J. Frenk. Health inequalities and social group differences: what should we measure? *Bulletin of the World Health Organization*, 77 (7):537, 1999.

National Records of Scotland. Health of Scotland's population - mortality rates. https://www2.gov.scot/Topics/Statistics/Browse/Health/TrendMortalityRates, 2016a. Accessed: 03-04-2020.

National Records of Scotland. The Scottish health survey 2016 edition. https://www2.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey/Publications, 2016b. Accessed: 03-04-2020.

P. Neal and T. Kypraios. Exact Bayesian inference via data augmentation. *Statistics and Computing*, 25(2):333–347, 2015.

NHS Health Scotland. Health inequalities: What are they? How do we reduce them? http://www.healthscotland.scot/media/1086/health-inequalities-what-are-they-how-do-we-reduce-them-mar16.pdf, 2015. Accessed: 26-9-2019.

A. Oliver. *Why care about health inequality?* Office of Health Economics, 2001. ISBN 1899040811.

F. C. Pampel, P. M. Krueger, and J. T. Denney. Socioeconomic disparities in health behaviors. *Annual Review of Sociology*, 36:349–370, 2010.

R. A. Pleasants, I. L. Riley, and D. M. Mannino. Defining and targeting health disparities in chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 11:2475, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

S. Reis, T. Liska, S. Steinle, E. Carnell, D. Leaver, E. Robert, M. Vieno, R. Beck, and U. Dragosits. *UK gridded population 2011 based on Census 2011 and Land Cover Map 2015*. 2017. doi: 10.5285/61f10c74-8c2c-4637-a274-5fa9b2e5ce44. URL https://doi.org/10.5285/61f10c74-8c2c-4637-a274-5fa9b2e5ce44. NERC Environmental Information Data Centre.

G. Roberts, A. Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

E. C. Rodrigues and R. Assunção. Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, 109:88–102, 2012.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

A. Rushworth, D. Lee, and R. Mitchell. 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, 2014.

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

Scheeres, Annaka. Kriging: Spatial interpolation in desktop GIS. https://www.azavea.com/blog/2016/09/12/kriging-spatial-interpolation-desktop-gis/, 2016. Accessed: 25-02-2020.

Scottish Government. Open access to Scotland's official statistics. https://statistics.gov.scot/home, 2019. Accessed: 26-11-2019.

V. Seaman. Article ii. *The Medical Repository of Original Essays and Intelligence, Relative to Physic, Surgery, Chemistry, and Natural History (1797-1800)*, 1(3):315, 1798.

N. J. Shelton. Regional risk factors for health inequalities in Scotland and England and the "Scottish effect". *Social science & Medicine*, 69(5):761–767, 2009.

G. D. Smith, M. Bartley, and D. Blane. The Black report on socioeconomic inequalities in health 10 years on. *BMJ: British Medical Journal*, 301(6748):373, 1990.

J. Snow. *On the mode of communication of cholera*. John Churchill, 2 edition, 1855.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

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

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

B. M. Taylor, R. Andrade-Pacheco, and H. J. Sturrock. Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1125–1150, 2018.

The Lancet Respiratory Medicine. Health inequality: a major driver of respiratory disease. *The Lancet. Respiratory Medicine*, 5(4):235, 2017.

B. Turner, P. Sederberg, S. Brown, and S. M. A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, 18:368–384, 2013.

M. Ugarte, J. Etxeberria, T. Goicoa, and E. Ardanaz. Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990–2005). *Cancer Epidemiology*, 36(3):254–262, 2012.

M. D. Ugarte, A. Adin, and T. Goicoa. Two-level spatially structured models in spatio-temporal disease mapping. *Statistical Methods in Medical Research*, 25(4):1080–1100, 2016.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8 (2):158–183, 2007.

J. Wakefield and R. Salway. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):119–137, 2001. ISSN 1467-985X. doi: 10.1111/1467-985X.00191. URL http://dx.doi.org/10.1111/1467-985X.00191.

D. Walsh, N. Bendel, R. Jones, and P. Hanlon. It's not 'just deprivation': why do equally deprived UK cities experience different health outcomes? *Public Health*, 124 (9):487–495, 2010.

D. Walsh, G. McCartney, C. Collins, M. Taulbut, and G. D. Batty. History, politics and vulerability: explaining excess mortality in Scotland and Glasgow. https://www.gcph.co.uk/assets/0000/5988/Excess_mortality_final_report_with_appendices.pdf, 2016. Accessed: 2-10-2019.

World Health Organization. Social determinants of health. https://www.who.int/social_determinants/thecommission/finalreport/key_concepts/en/, 2013. Accessed: 25-9-2019.

World Health Organization and others. Risk factors for chronic respiratory diseases. *Global Surveillance, Prevention and Control of Chronic Respiratory Diseases: a*

Comprehensive Approach. Geneva, Switzerland: World Health Organization, pages 37–55, 2007.