Estimating the changes in health inequalities across Scotland over time

Eilidh Jack

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy

July 2019

Declaration

I, Eilidh Jack, declare that this thesis titled, 'Estimating the changes in health inequalities across Scotland over time.' and the work presented in it are my own. I confirm that where I have consulted the published work of others, this is always clearly attributed.

The work presented in Chapter 4 has been published in the Journal of the Royal Statistical Society - Series A (JRSSA) with the title 'Estimating the changing nature of Scotland's health inequalities using a multivariate spatio-temporal model' (Volume 182(3), p1061-1080), and is jointly authored by Prof Duncan Lee and Dr Nema Dean. I delivered an invited talk on this work at the GEOMED conference in Porto, Portugal in 2017 and a topic contributed talk at the JSM conference in Vancouver, Canada in 2018.

There is little success where there is little laughter.

— A. Carnegie

Abstract

Health inequalities are the unfair and avoidable differences in people's health between different social groups. These inequalities have a huge impact on people's lives, particularly those who live at the poorer end of the socio-economic spectrum, as they result in prolonged ill health and shorter lives. Much of the existing research into health inequalities in Scotland lacks analysis at the small area scale. The work in this thesis aims to fill that gap by estimating health inequalities in Scotland over time at the small area level used for data collection, which are known as intermediate geographies (IGs), as well as between Scotland's 14 regional health boards, which are responsible for the protection and improvement of their populations' health. This thesis utilises conditional autoregressive (CAR) models which are the most common modelling approach for areal unit data. The first model proposed aims to estimate inequalities in risk of coronary heart disease from 2003 to 2012 across Scotland. However, focusing on a single disease gives an incomplete picture of the overall inequality in population health. Therefore, the second model proposed is a novel multivariate spatio-temporal model for quantifying health inequalities in Scotland across multiple diseases, which will enable us to better understand how these inequalities vary and correlate across diseases and how they have changed over time. This methodology is applied to hospital admissions data for cerebrovascular disease, coronary heart disease and respiratory disease, three of the leading causes of death, from 2003 to 2012 across Scotland. Finally, it was identified that a common problem in areal unit data of this type is changes to boundaries which occur during the time period for which data are available. This occured in Scotland when the IG boundaries were redrawn after the 2011 census. The final

piece of work in this thesis aims to address the problem of spatial misalignment by proposing a multiple imputation approach which utilises a common latent spatial grid. This approach is applied to data containing hospital admissions for respiratory disease for the years 2006 - 2016 for the health board Greater Glasgow and Clyde, where the data from 2013-2016 are reported on the areas with redrawn boundaries. Overall, it was found that there are still considerable health inequalities in Scotland at both the small area level and between Scotland's health boards. Although these inequalities appear to be decreasing over time for cerebrovascular and coronary heart disease, they are increasing for respiratory disease. In particular, the risk of most areas which were estimated to have a high risk of respiratory disease at the start of the time period are increasing at a higher rate than areas with low risk. It was also found that areas which experience high risk of one disease tend to experience high risk of all three diseases studied here. This highlights the issue that Scotland is facing and that more needs to be done to target the areas which are experiencing high risk of disease across multiple diseases.

Acknowledgements

Firstly, I would like to thank my supervisors Prof Duncan Lee and Dr Nema Dean. I have been lucky enough to have had an almost entirely positive experience as a PhD student and this is in large part due to the excellent supervision I have recieved from you both. From the start you have been nothing but supportive and inspiring, and have helped me in ways that go beyond your duties as PhD supervisors.

I would also like to thank the Carnegie Trust for their generous funding.

Thank you to all of my friends who have kept me smiling throughout. My fellow PhD students, thank you for your support and help with all things PhD and beyond, and for always being there for a chat when needed. In particular, a huge thank you to Suzy. I feel so lucky to have had you by my side throughout this experience. You were always at hand with a cup of tea after supervisor meetings to help me think through my problems or distract me from them when necessary. You have made the hard times bearable and the good times so much better.

Thank you to my family, in particular, my parents for your support and guidance throughout my life. I would not be in the position I am today without you. Thank you to my Gran and Gran'pa for all that you have done for me. You are, and will always be, an inspiration to me in all aspects of my life. To Ross, for your unwavering belief in me and your constant reassurance when things seemed beyond my reach. For helping to celebrate the successes and making the failures seem less significant, you have helped me more than you know.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Disease risk is not constant over space and time and is often impacted by exposure to risk inducing behaviour such as consumption of alcohol. Poverty, and more generally deprivation, are major factors in the spatial variation observed in the risk of disease, with more highly deprived areas usually exhibiting elevated levels of disease risk (McCartney, 2012). This difference in disease risk between social groups and population areas is known as a health inequality (or inequity). Importantly, health inequalities refer to the unfair and avoidable differences in people's health, and are based largely on socio-economic factors such as income, wealth and education. They are fundamentally driven and shaped by economics, social policy and politics, which in turn lead to an unequal distribution of income, power and wealth. On a global level the world's poorest people tend to have the worst health, and so health inequalities exist between countries. For example, the average life expectancy in Japan is 83.7, compared to 50.1 in Sierra Leone (World Health Organization, 2016).

Health inequalities also exist to a large extent within countries, and are seen in countries of all ranges of incomes. There is evidence which shows that individuals who live at the poorer end of the socio-economic spectrum exhibit poorer health regardless of the wealth of the country (World Health Organization, 2008). The first major report on health inequality in the UK was The Black Report (Black et al., 1982), which was commissioned by the Labour Government in 1977. The report showed the unequal distribution of ill-health and death across the UK, and concluded that these

inequalities were due mainly to social inequalities affecting health. The Acheson Report (Acheson, 1998) confirmed findings from The Black Report that the '*weight of scientific evidence supports a socio-economic explanation of health inequalities*'. Following the publication of these reports, it became widely recognised that social class had a strong bearing on life expectancy. This is illustrated in Table 2.1 of Bartley (2016), which shows standardised mortality ratios (SMR) by Registrar-General's Social Class (RGSC) in men aged 15-64, with an SMR for class I (professional) of 66 compared to 189 for class V (unskilled manual) in 1991. More recently, The Marmot Review (Marmot, 2010) was published in 2010, and its key policy objectives focus on the social determinants of health.

Reports such as these have aided the development of several social models which differ from statistical models in that the aim is to explain the behaviours of the people involved in a certain event or activity. Three of the most common models of explanation for health inequalities, described in detail in Bartley (2016), are behavioural, material and psycho-social. Very briefly, behavioural and cultural explanations refer to the existence of health inequality due to differences in life-style between social groups, most notably, smoking, exercising for leisure, and quantity of fats, sugars and salt in the diet. The psycho-social model identifies the psycho-social risk factors which impact health, including social support, autonomy at work and the balance between home and work (Hemmingway and Marmot, 1999). The materialistic model recognises the significance of '*the...diffuse consequences of the class structure: poverty, work conditions...and deprivation in its various forms in the home and immediate environment, at work, in education and the upbringing of children and more generally in family and social life*' (Black et al., 1982). Much of the literature attempts to use a combination of these three models when attempting to explain the causes of health inequality.

In this thesis, I focus on estimating health inequalities in Scotland for several reasons. First of all, Scotland has very poor health for a European country, with the lowest and most slowly improving life expectancy compared to all other western European countries (Walsh et al., 2016). This can be seen in Figure 1.1 which shows the male and female life expectancy for Scotland in blue compared to 19 other western

European countries in red from 1851 - 2005. Scotland also has the widest health inequalities in western Europe (Popham and Boyle, 2011). For example, in 2015 it was estimated that in the most affluent areas in Scotland, men experience 23.8 more years of good health, and women 22.6, compared to those living in the most deprived areas (NHS Health Scotland, 2015). Although there was a major reduction in health inequalities in Scotland between 1920 and 1970 (Beeston et al., 2013), clearly gaps still exist between Scotland's most affluent and most deprived areas, and in some cases these gaps are still widening. This is illustrated in Figure 1.2, which shows the average life expectancy of males living in the vicinity of each stop of the Glasgow subway line (Glasgow Open Data, 2010). The 15 stops on the Glasgow subway are distributed over a 10km circuit which takes only 24 minutes to complete, and so the differences in the life expectancies between some stops are substantial for what is a relatively small geographical area. For example, at Hillhead the average life expectancy is 80, which is 14 years higher than at Govan, only a 6 minute subway journey away. On an individual level this is a tale of human tragedy, with too many Scots experiencing poor health and dying prematurely as a direct result of health inequalities. These inequalities in health also have huge economic repercussions for Scotland and place an enormous burden on the NHS. For example, it has been estimated that if the death rate across Scotland fell to the level of the least deprived areas, the economic benefit could exceed £20billion (Audit Scotland, 2012).

Of all the problems facing Scotland, high mortality rates and gaping health inequalities are clearly of huge political and social importance, and reducing these inequalities has been a priority for the Scottish Government for many years. In 2007 they established a Ministerial Task Force for Health Inequalities whose aim was '*to identify and prioritise practical actions to reduce the most significant and widening health inequalities*' (The Scottish Government, 2008). There have been many reports by official bodies such as The Scottish Government, NHS Scotland and Audit Scotland (The Scottish Government, 2008, NHS Health Scotland, 2015, Audit Scotland, 2012) on Scotland's health inequalities, with a focus on ways to reduce these in the future.

Given the importance and extent of health inequalities across the globe, and

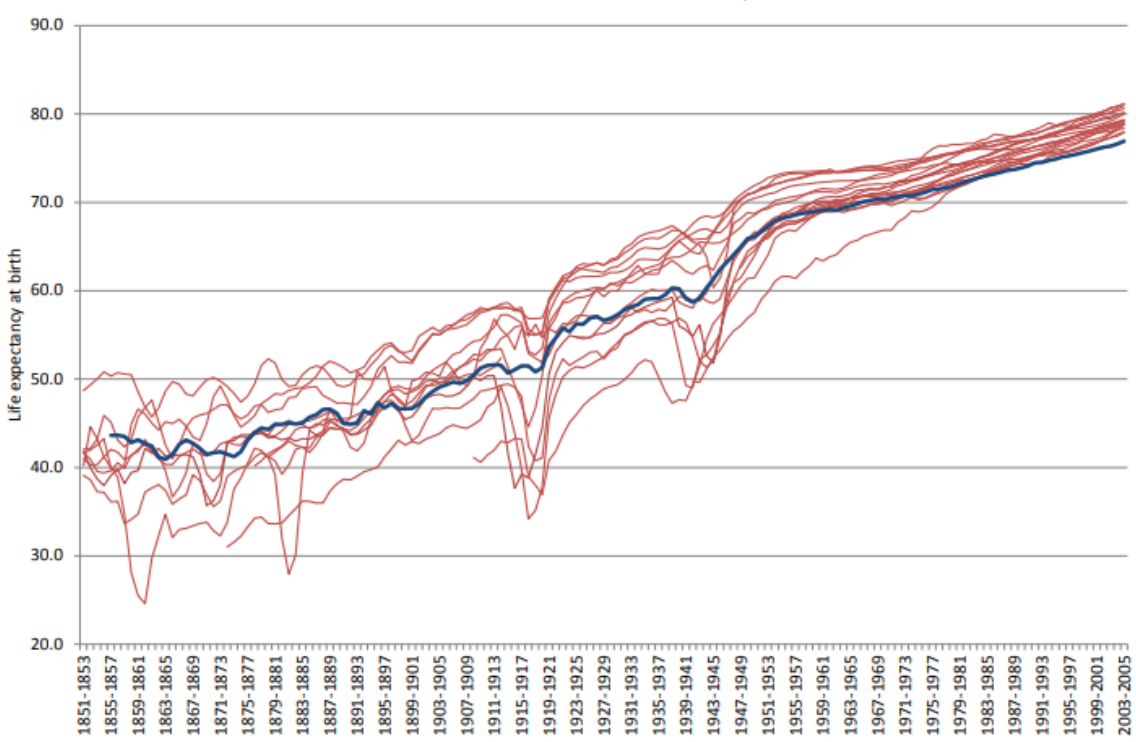

Figure 1.1: Male and female life expectancy for Scotland (blue) compared to 19 other Western European Countries (red) from 1851-2005 Plot source: Walsh et al. (2016). Data source: Human Mortality Database

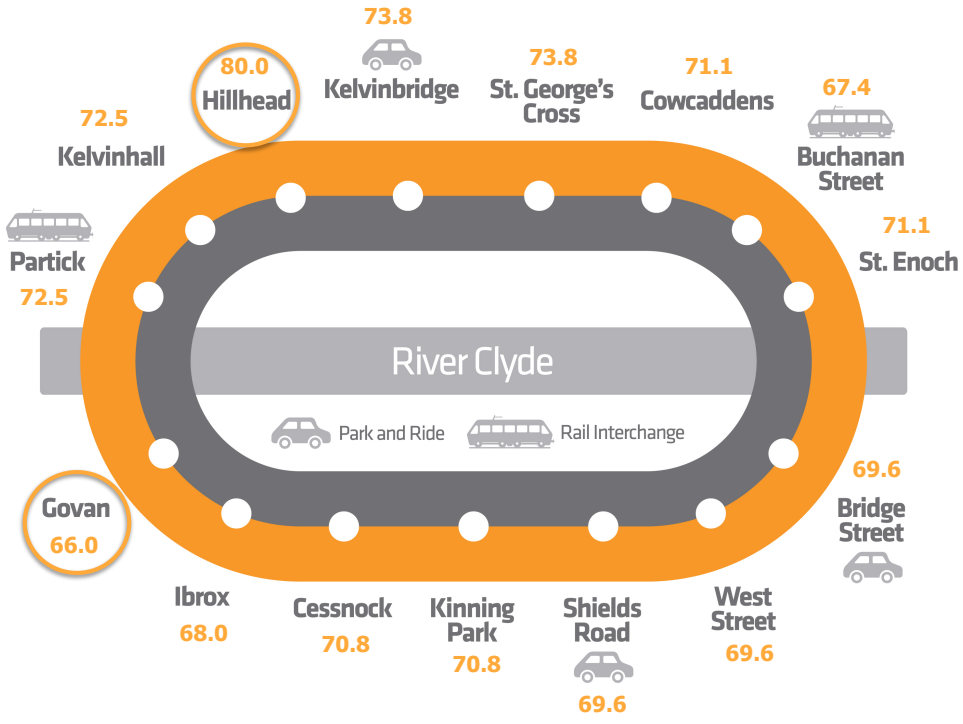

Figure 1.2: Map of the Glasgow subway with male life expectancies for each stop.
Source of subway map image: SPT Subway Maps & Stations page.

more particularly in Scotland, there has been extensive research in this area. For example, Taulbut et al. (2014) compared West Central Scotland with other post-industrial regions of Europe, and Leyland et al. (2007) examined patterns in, and causes of, inequalities for regions of Scotland. Comparisons between other European countries and Scotland were the focus of Walsh et al. (2016), but they concentrated on explaining Scotland's, and particularly Glasgow's, excess mortality. However, this research collectively lacks an in-depth analysis of health inequalities for multiple diseases at the small area scale in Scotland, which is the focus of this thesis.

In this thesis, I estimate health inequalities within Scotland at a small area level to allow for a more in-depth understanding of where in Scotland these inequalities exist and how they are changing over time. This general area of spatial epidemiology is known as disease mapping and has been growing in popularity because of its potential usefulness in regional health planning, disease intervention and allocating health funding. It allows for the construction of smoothed spatial maps of disease risk and the assessment of possible determinants of diseases. Most disease mapping approaches tend to utilise data collected on non-overlapping areal units, such as census tracts or electoral wards, as patient-level data cannot be made publicly available due to patient confidentiality. However, studies at this small area level are often of more use to health authorities as they are interested in risk levels across the whole population.

The study region considered in this thesis is Scotland, which is the UK's northernmost country. The population of Scotland is around 5.4 million, and it consists of a variety of different landscapes including highly populated cities, mountainous wilderness and several small and sparsely populated islands. The small area units for which data are available are known as intermediate geographies (IGs) (http://www.gov.scot/Publications/2005/02/20732/53083), and many publicly available datasets are published at this level. There are a total of 1235 IGs which, on average, contain 4000 household residents. The geographical size of these IGs varies widely and is dependent on the population density of the underlying area. Figure 1.3 shows a map of the IGs in Scotland and it can be seen that in the densely populated 'central belt' of Scotland, the geographical area of each IG is much smaller than the

sparsely populated areas to the north. As well as quantifying health inequalities between IGs in Scotland, we are also interested in estimating the health inequalities that exist between Scotland's 14 regional NHS health boards, which are responsible for the protection and improvement of their populations' health and for the delivery of frontline healthcare services. The 14 HBs are a range of sizes, both geographically and in terms of population, which can be seen from Figure 1.4. The estimated population and number of IGs in each HB is provided in Table 1.1. The HBs with the smallest populations are two of the island boards, Orkney and Western Isles, with 2012 population estimates of 21,530 and 23,210 respectively. Greater Glasgow and Clyde, and Lothian are the most populated HBs despite their relatively small geographical area, with 2012 estimates of 1,217,025 and 843,733 respectively. This is due to both of Scotland's largest cities being situated in these HBs, Glasgow and Edinburgh respectively, which can be seen from Figure 1.4. This shows that a large land area does not imply a large population and instead the number of IGs is driven by the population size rather than geographical size, with Greater Glasgow and Clyde having the most IGs of all the HBs. In 2011/12 The Scottish Government allocated around £170million to the health boards to address health inequalities, giving them direct responsibility for tackling this problem (Audit Scotland, 2012). A key focus of the research in this thesis is therefore to estimate the scale of health inequality in disease risk between these health boards and investigate how this is changing over time.

The first aim of this thesis will be to develop a single disease spatio-temporal model to estimate health inequality at both the small area IG and larger HB level. This methodology will then be extended into a multivariate setting using a novel multivariate spatio-temporal model which will allow for health inequalities in multiple diseases to be estimated in one model. This will fill a gap in the literature and provide a clearer picture of overall health inequality in Scotland and how it has changed over time. The methodology is applied to hospital admissions data for three of Scotland's biggest killers (Scotpho, 2016), namely, cerebrovascular disease, coronary heart disease and respiratory disease from 2003 to 2012 across Scotland. Finally, it is not uncommon for the boundaries associated with areal unit data to change over

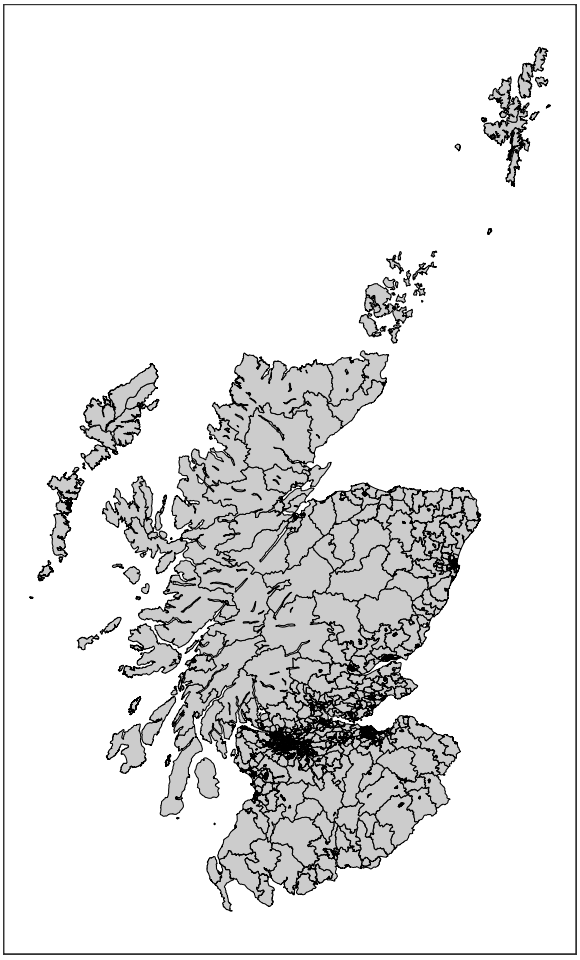

Figure 1.3: Map of the intermediate geographies in Scotland.

Table 1.1: Information on Scotland's 14 health boards.

Health board	Code	Estimated population (2012)	# of IGs
Ayrshire and Arran	A	373,189	92
Borders	B	113,707	29
Fife	F	366,219	103
Greater Glasgow and Clyde	G	1,217,025	272
Highland	H	319,811	76
Lanarkshire	L	572,520	137
Grampian	N	573,420	128
Orkney	R	21,530	6
Lothian	S	842,733	177
Tayside	T	411,749	90
Forth Valley	V	299,099	74
Western Isles	W	27,560	9
Dumfries and Galloway	Y	150,828	35
Shetland	Z	23,210	7

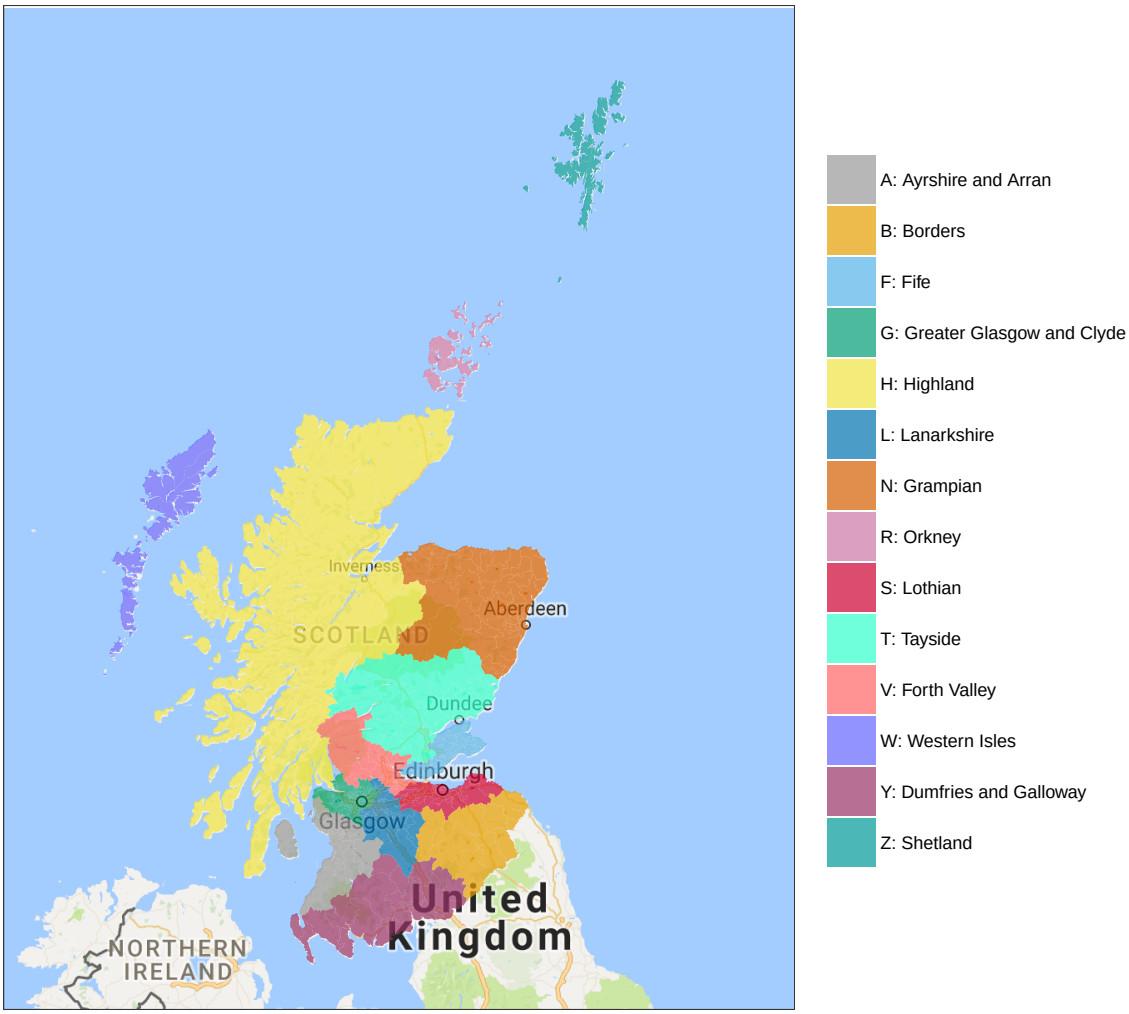

Figure 1.4: Map of the NHS health boards in Scotland.

the time period for which data are available. After the 2011 population census, the Scottish Government decided to redraw the boundaries of the data zones which are the key geography for small area statistics in Scotland and are used to create the intermediate geographies that are used throughout this thesis. The redrawn data zones were released in 2014 along with new boundaries for intermediate geographies. Statistically, this poses a challenge since using data from before and after this change would lead to non-comparable inference due to spatial misalignment of the IG data. The final piece of work in this thesis aims to address this problem by using a multiple imputation approach to undertake inference on a common grid for both sets of IGs, thus producing comparable inference over time. This approach will then be applied to data containing hospital admissions for respiratory disease for the years 2006 - 2016 for the Greater Glasgow and Clyde health board, where the data from 2013 - 2016 are reported on the redrawn IGs.

The remainder of this thesis is organised as follows. Chapter 2 provides an overview of some of the existing methodology which will be used in this thesis and provides a review of the relevant literature, with particular focus on spatial, spatio-temporal and multivariate spatial/spatio-temporal methodology. In Chapter 3 a spatio-temporal hierarchical Bayesian model is developed and applied to data containing coronary heart disease hospital admissions for the years 2003 to 2012. The main aim of this chapter is to quantify health inequalities between both Scotland's IGs and HBs and investigate how these have changed over time. The model presented in Chapter 3 is then extended in Chapter 4 to provide a novel multivariate spatio-temporal model to allow for health inequalities to be compared across three diseases. Although estimating health inequalities between IGs and HBs within each disease is still of interest, another key aim is to investigate if there any differences in these health inequalities across the three diseases. In Chapter 5 a multiple imputation approach is developed to undertake inference on a common grid which will allow for comparable inference over time when the spatial boundaries associated with areal unit data change over time. The main aim of this chapter is to quantify health inequalities in respiratory disease in Greater Glasgow and Clyde and investigate how these have changed over a time period with a change in areal unit boundaries. Fi-

nally, Chapter 6 provides a summary of the key findings of this thesis and possible future work.

Chapter 2

Literature Review

This chapter outlines the statistical methodology which is used throughout this thesis, as well as giving an overview of the exisiting literature within these areas of statistics. Section 2.1 introduces disease mapping which is an area of spatial epidemiology which allows for disease risk to be estimated. Section 2.2 gives a brief overview of Bayesian statistics, which is the statistical framework utlilised thorughout this thesis. Section 2.3 introduces generalised linear models (GLMs), with particular focus on GLMs for count data. Some aspects of geostatistical modelling which are employed in this thesis are described in Section 2.4. Sections 2.5 and 2.6 explore some of the existing literature in spatial and spatio-temporal modelling which form the basis of Chapter 3. Sections 2.7 and 2.8 introduce existing methodology for multivariate spatial and multivariate spatio-temporal models. Finally, Section 2.9 explores existing literature used for continuous inference on aggregated data.

2.1 Disease Mapping

The risk of a particular disease can often vary over space, and geographic patterns of disease can be attributed to many risk factors such as differences in environmental exposures and in the behaviours of the inhabitants of different areas. As previously mentioned, poverty and more generally deprivation are major factors in the spatial variation observed in the risk of disease, with higher levels of disease risk generally observed in more highly deprived areas (McCartney, 2012). In order to assess the

extent and pattern of differences in disease risk over space, estimates of risk are presented on a disease map which are computed by partitioning the study region into n contiguous small areas and computing and then mapping the disease risk in each area. The most popular way of illustrating differences in disease risk is via a choropleth map where areas are shaded on a scale relating to disease risk. This area of spatial epidemiology is known as disease mapping and has growing popularity because of usefulness in regional health planning, disease intervention and the allocation of health funding.

Most disease mapping approaches tend to utilise data collected on non-overlapping areal units, such as census tracts or electoral wards, as patient level data cannot be made publicly available due to patient confidentiality. In general, areal unit data are data collected over a study region $\boldsymbol{\mathcal{A}}$ which is partitioned into n contiguous small areas $\{\mathcal{A}_1, ..., \mathcal{A}_n\}$. For each of these areas, a response is observed to give $\boldsymbol{Y} = (Y_1, ..., Y_n)$.

The naive approach would be to model disease risk using the disease counts $\boldsymbol{Y} = (Y_1, ..., Y_n)$, however this ignores the substantial differences in the population sizes and demographics which could account for some of the differences in the disease counts between areas. Due to these differences in the demographic structures of each area, expected counts e_i for area i are usually calculated using indirect standardisation. This is done by splitting each area into $j = (1, ..., J)$ strata (for example ten year age bands split by gender). The expected counts are then calculated by

$$e_i = \sum_{strata\, j} r_j N_{ji}, \tag{2.1}$$

where r_j is the risk of disease for strata j for the entire study population and N_{ji} is the population size for strata j in area i. Based on these counts the simplest measure of disease risk is the standardised incidence ratio (SIR), which is the ratio of the observed counts and the expected counts of disease cases for each areal unit, $SIR_i = Y_i/e_i$. Values of SIR greater than 1 represent elevated levels of disease risk, and values less than 1 correspond to decreased levels of disease risk, for example, an SIR of 1.2 corresponds to a 20% increase in risk for that area. However, the SIR can give unstable and uninformative estimates when the data includes areas where

the expected numbers of disease cases are small. In order to overcome this issue a hierarchical Bayesian modelling approach is typically adopted to estimate the risks, using a combination of covariate information and a set of spatially varying random effects. These random effects borrow strength from neighbouring areas which reduces the chance of excesses in risk occurring randomly.

Although the geographical location of each areal unit is of direct interest, there is no reason to believe that location itself would affect the risk of a certain disease. Generally, the spatial location is a proxy measure of differences in the attributes of an area such as social differences, e.g. consumption of alcohol, or differences in physical or environmental geography, e.g. temperature or air quality. Some of these factors may be known in advance to have an effect on the risk of disease and, if data are available, can then be included in the model as a covariate. For example, a major factor in health inequality is socio-economic deprivation, and so a measure of this is often included in disease mapping models.

Spatial random effects are typically included in a model to account for any residual spatial correlation left in the data after the covariate effects have been removed, and can be seen as a surrogate for missing covariates (which are either unknown or unable to be measured) that are correlated with location. Prior models are discussed for these random effects in Section 2.5, but first an introduction to Bayesian modelling is given.

2.2 Bayesian statistics

2.2.1 Introduction

The notion of probability in a frequentist setting is attached only to repeatable random events. Probabilities are never assigned to any fixed and unknown quantities. So the observed data $\boldsymbol{Y} = (Y_1, ..., Y_n)$ are treated as a repeatable random sample and the underlying parameters of interest, say $\boldsymbol{\theta}$, are assumed to remain fixed. Maximum likelihood methods are typically used to find parameter estimates based on data collected from a sample of the population and often large sample properties or asymptotic approximations are needed. For details see Garthwaite et al. (2006).

Compare this to a Bayesian framework where the parameters, $\boldsymbol{\theta}$, are treated as

random variables. In other words, the uncertainty around the unknown parameters of interest is represented probabilistically. This approach has origins from the mid-18th century in Bayes' Theorem which was developed by Thomas Bayes (Bayes, 1763) and is defined as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{Y})}. \tag{2.2}$$

Here $f(\boldsymbol{\theta}|\boldsymbol{Y})$ denotes the posterior distribution of the parameters given the data, $f(\boldsymbol{Y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta})$ denotes the data likelihood, $f(\boldsymbol{\theta})$ denotes the prior distribution which expresses our beliefs about the parameters before we see the data, and $f(\boldsymbol{Y})$ is the normalising constant, which typically we don't need to calculate explicitly. Therefore we can rewrite Equation 2.2 up to a constant of proportionality as

$$f(\boldsymbol{\theta}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \tag{2.3}$$

The posterior distribution tells us everything we need to know about the parameters. However, we may want to find summaries of this distribution. For example, central point estimates can be found by taking the mean or median of the posterior distribution. Unlike frequentist statistics, in Bayesian statistics we can make probability statements about model parameters, and uncertainty in point estimates can be estimated using interval estimates. For example, a credible interval is a range of values that the parameter can take with a particular probability (say 0.95) and is found by computing the end points of an interval that correspond with specified percentiles of the posterior distribution (e.g. the 2.5th percentile and the 97.5th percentile).

Although Bayes' theorem was proposed by Thomas Bayes in the mid-18th century, the adjective 'Bayesian' was not part of statisical vocabularly until far more recently. Many inferential methods which were based directly on the use of Bayes Theroem were commonly referred to as 'inverse probability' up until the middle of the twentieth century (Flenberg, 2006) and frequentist-based solutions were often not available for complicated problems. However, from the 1950's onwards there were huge advancements in computers and computing power which allowed solutions of complex models to be found in a Bayesian framework, as it is now known. Today the choice

between frequentist and Bayesian approaches is the subject of much debate among statisticians. However, although the philosophical differences underlying both lead to different approaches they can often yield similar results. The problems tackled in this thesis will be approached using Bayesian hierarchical models as these can help in understanding multiparameter problems and allow for models to be developed naturally as levels of complexity are added. The following sections will summarise some key aspects of Bayesian modelling.

2.2.2 Prior distributions

One of the biggest criticisms of the Bayesian paradigm is the use of prior distributions. A frequentist will argue that choosing a very informative prior will necessarily bias your results and if no information is known about the parameter then the choice of the prior can be problematic. This section will outline some of the common types of prior distribution used in practice.

Priors should be chosen before the data has been seen and often there may be some prior information available from previous studies. In this case, the information can be used to formulate an informative prior. Conversely, if nothing is known about the parameter, then a weakly informative prior could be assigned. In this case, the posterior distribution will be dominated by the likelihood function and estimated mainly from the data. An example of a weakly informative prior for real valued parameters would be a Gaussian distribution with a large variance, e.g. $\theta_j \sim \mathrm{N}(0, 10000)$.

Another important concept is the use of proper versus improper priors. If a prior distribution does not integrate to a finite number when integrated over its range space, it is an improper prior. For example, a uniform prior distribution on the real line, $f(\theta) \propto 1$, for $-\infty < \theta < \infty$. Although improper priors can lead to proper posteriors, this is not always the case and care must be taken to check that the posterior distribution is proper, as inference cannot be made with improper posterior distributions.

Often priors are chosen through convenience, to make inference more easily achieved. If a conjugate prior to the likelihood is chosen then the posterior will be of the same distributional form as the prior and so will come from a standard distributional fam-

ily. If this is the case, evaluation is more easily achieved as we are sampling from a known distribution. Consider data $Y_i \sim \text{Poisson}(\lambda)$ for $i = 1, ..., n$, an example of a conjugate prior distribution for the parameter λ would be $\lambda \sim \text{Gamma}(\alpha, \beta)$. Here the posterior distribution will be of the same form as the prior, i.e. $f(\lambda|\boldsymbol{Y}) \propto \text{Gamma}(\alpha + \sum_{i=1}^{n} Y_i, \beta + n)$.

In this thesis, I will make use of both conjugate and weakly informative priors as well as more informative priors which are based on our prior beliefs about the data.

2.2.3 Inference

Once the prior distribution has been chosen and an appropriate posterior distribution has been derived, Bayesian inference relies on the ability to compute this posterior distribution. In simple cases, closed form solutions of the posterior distribution may exist or Monte Carlo approximation techniques can be used to provide a sample from the distribution of interest. However, it is not always possible to sample directly from a target distribution so instead more complex methods are used which generally rely on approximating the posterior distribution in some way.

The most common method used to do this is a general class of algorithms called Markov chain Monte Carlo (McMC), which provides dependent samples from the 'target' posterior distribution. Generally, a Markov chain with equilibrium distribution equal to the target distribution is constructed, draws are then simulated until the Markov chain has converged, i.e. the current state of the Markov chain is approximately a draw from the posterior distribution. There are two main ways to achieve McMC simulation, Gibbs sampling (Geman and Geman, 1984) and the Metropolis-Hastings algorithm (Hastings, 1970), which are both used in this thesis and are described in the following sections.

2.2.3.1 Gibbs sampling

Suppose the parameter $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_p)$, and we want to sample from the joint posterior distribution $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_p|\boldsymbol{Y})$. Then if the full conditional distribution for each $\boldsymbol{\theta}_i$, $f(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{Y})$, where $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, ..., \boldsymbol{\theta}_p)$, is known, then we can use Gibbs sampling to sample from this distribution. Often conjugate

priors, described in Section 2.2.2 are specified in order to use Gibbs sampling to estimate the posterior distribution. The general algorithm for Gibbs sampling the parameter $\boldsymbol{\theta}$ is as follows.

1. Choose an initial value, $\boldsymbol{\theta}^{(0)}$ to start the chain. For $t = 1, 2, ...$
2. Let the current state of the Markov chain be $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, ..., \boldsymbol{\theta}_p^{(t)})$, then at time $t + 1$:
   1. draw $\boldsymbol{\theta}_1^{(t+1)}$ from $f(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, ..., \boldsymbol{\theta}_p^{(t)})$
   2. draw $\boldsymbol{\theta}_2^{(t+1)}$ from $f(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, ..., \boldsymbol{\theta}_p^{(t)})$

   ...

   p. draw $\boldsymbol{\theta}_p^{(t+1)}$ from $f(\boldsymbol{\theta}_p | \boldsymbol{\theta}_1^{(t+1)}, ..., \boldsymbol{\theta}_{p-1}^{(t+1)})$
3. Increase t by 1 and repeat the above steps until we have the desired number of draws.

2.2.3.2 Metropolis-Hastings algorithm

If, however, some or all of the full conditionals, $f(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \boldsymbol{Y})$, are not from a known family of distributions then the Metropolis-Hastings algorithm can be used to sample from these distributions. The general recipe is as follows for a single block $\boldsymbol{\theta}_i$.

1. Initialise $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$. At step $t = 1, 2, ...$

   1 Given the current point, $\boldsymbol{\theta}_i^{(t)}$ simulate $\boldsymbol{\theta}_i^*$ from proposal distribution, $g_t(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i^{(t)})$.

   2 Evaluate the acceptance ratio

$$r = \frac{f(\boldsymbol{\theta}_i^* | \boldsymbol{Y}) g_t(\boldsymbol{\theta}_i^{(t)} | \boldsymbol{\theta}_i^*)}{f(\boldsymbol{\theta}_i^{(t)} | \boldsymbol{Y}) g_t(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i^{(t)})}. \tag{2.4}$$

   If the proposal distribution is symmetric, this can be reduced to the Metropolis algorithm:

$$r = \frac{f(\boldsymbol{\theta}_i^* | \boldsymbol{Y})}{f(\boldsymbol{\theta}_i^{(t)} | \boldsymbol{Y})}. \tag{2.5}$$

   3 Pick $U \sim \text{Uniform}(0, 1)$. If $U < r$, take $\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^*$, otherwise $\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t)}$.

The proposal distribution $g_t(\boldsymbol{\theta}_i^*|\boldsymbol{\theta}_i^{(t)})$ is used to propose new values for the parameters based on their current values and can be chosen to be anything as long as it samples from a valid range, for example a symmetric proposal distribution could be N $\sim$ $(\boldsymbol{\theta}_i^{(t)}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}_i^{(t)}$ is the current sampled value of $\boldsymbol{\theta}_i$. It is this which determines how likely the new value is to be accepted. For example, if the proposed value is close to the current value (proposal distribution has small variance) then the proposed value is more likely to be accepted, whereas if the proposed value is very different to the current value (proposal distribution has large variance) it is less likely to be accepted. In general, the number of proposals which are accepted should not be too high or too low. For example, Roberts and Rosenthal (2001) propose that for random walk Metropolis on smooth densities, any acceptance rate between 0.1 and 0.4 should perform close to optimal. When the jumps made are very large each time, the acceptance rate will be too low and the Markov chain could become stuck on one value for long periods of time. Conversely, if the jumps made are too small, the acceptance rate will be too high and the sampler will take too long to explore the parameter space. In this thesis, this will be controlled by updating the proposal variance every 100 iterations if the acceptance rate is deemed too high (greater than 0.4) or too low (less than 0.1).

2.2.3.3 Model convergence

In practice, often a combination of both Gibbs sampling and Metropolis-Hasting steps are implemented to estimate posterior distributions for the parameters of interest in a hierarchical Bayesian model. The number of draws that the sampler needs to run for before the Markov chain has converged will depend on the complexity of the underlying statistical model. In practice, a trial and error approach tends to be used with convergence checked after the chain has been sampled for M draws. This is a crucial stage of the process as inference is only valid after our Markov chain has converged to our target distribution. We want to ensure that there are enough McMC samples from high posterior density regions of our target distribution, i.e. that our Markov chain has achieved stationarity and our Markov chain is therefore ergodic. This then allows us to calculate the quantities of interest from our draws, ignoring the

dependence between draws. We also want to ensure that the sampler moves between the separate regions of high probability. This is known as mixing. One method is to examine trace plots of the McMC samples for individual parameters, which show the parameter values during the runtime of the chain and should show no trend. The time taken for the chain to reach convergence is known as the burn-in period, and these samples are discarded before inference is made as they are not representative samples from the target distribution. Another tool for checking model convergence is the Geweke diagnostic (Geweke, 1992), which is based on a test for equality of the means of the first and last part of a Markov chain (usually the first 10% and the last 50%). The test statistic is a standard Z-score meaning that values between (-1.96 and 1.96) are indicative of convergence.

Another problem with McMC sampling is that the draws will show within chain correlation since each draw is dependent on the previous value. Once convergence has been checked, the burn-in discarded and the chain is now of length M, these samples are not independent and therefore contain less information about the parameter than M independent samples would, i.e. the *effective* number of independent samples is far fewer than M. One suggested approach to tackle this is to *thin* the chain by only keeping every k^{th} simulation draw and discarding the rest. However it has been argued that thinning is very wasteful of information, often unnecessary and increases computation time (Link and Eaton, 2012). It can, however, be useful when computer storage is a problem. For example, when the number of parameters is very large. This is the case in this thesis and so thinning is used to produce samples closer to independent and reduce the number of samples which need to be stored.

One of the disadvantages of McMC techniques, particularly for complex problems, is that the algorithm can involve sampling a large number of parameters and so can take a long time to run before a representative sample from each posterior distribution has been drawn. One technique which can be used to speed up this process is blocking, which involves constructing the sampler in a way that several parameters are drawn at the same time. This should improve the efficiency of the sampler, particularly if the parameters are correlated. However, as you increase the number of parameters in the block and the dimensionality of the proposal distribution increases, it can be

easy to miss important parts of the parameter space and acceptance rates generally reduce.

2.3 Generalised linear models

Consider a linear regression model of the form:

$$\mathbb{E}(Y_i) = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.6}$$

where;

- $Y_i \sim N(\mu_i, \sigma^2)$, and are independent for $i = 1, ..., n$.
- $\boldsymbol{x}_i = (1, x_{i2}, ..., x_{ip})^\top$ is the ith row of the design matrix, $\boldsymbol{X}$, of known covariates.
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is the unknown parameter vector.

This can be generalised to allow for a response variable which follows a distribution other than normal, but which belongs to a very flexible class of distributions known as the exponential family of distributions. Consider the random variable Y whose probability distribution function (p.d.f) depends on the parameter θ. The distribution belongs to the exponential family if it can be written as:

$$f(y|\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{2.7}$$

or equivalently:

$$f(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)], \tag{2.8}$$

where $s(y) = \exp[d(y)]$ and $t(\theta) = \exp[c(\theta)]$.

Generalised linear models (GLMs) (Nelder and Wedderburn, 1972) include a relationship between the response and the linear component of the form:

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.9}$$

where g is a monotone, differentiable function called the link function which describes how the mean, $\mathbb{E}(Y_i) = \mu_i$, depends on the linear predictor. What is of interest in a

GLM are the parameters $(\beta_1, ..., \beta_p)$ that describe how the response depends on the explanatory variables.

2.3.1 Generalised linear models for count data

In this thesis, all the methodology is developed to model count data, such as numbers of hospital admissions. The standard model used to represent data such as this is the Poisson distribution. As previously discussed, due to differences in the demographic structures of each areal unit disease risk cannot be modelled purely on the disease counts, $\boldsymbol{Y} = (Y_1, ...Y_n)$, and so expected disease counts, $\boldsymbol{e} = (e_1, .., e_n)$, are calculated and now $\mu_i = e_i\theta_i$, where θ_i is the disease risk for area i. Since the Poisson distribution is a member of the exponential family of distributions, a generalised linear model can be used to model these data. The general form is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(e_i\theta_i), \quad i = 1, ..., n, \\ \log(\theta_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta}, \end{aligned} \tag{2.10}$$

where the link function is the natural log and θ_i is the risk of disease in area i and is on the same scale as the SIR defined previously. One of the key features of the Poisson distribution is that the mean is equal to the variance, so

$$\text{var}(Y) = \mathbb{E}(Y) = \mu. \tag{2.11}$$

However, often in practice we find that the variance is larger than the mean, which is known as overdispersion. One way to overcome this issue is to use a quasi-poisson distribution instead, which makes the less restrictive assumption that the variance is proportional to the mean (McCullagh and Nelder, 1989). Another potential issue with count data arises when working with data that contains an excess of zero counts, for example rare disease data. In this case, it has been suggested that the excess zeros are generated by a separate process than the counts and so a zero-inflated Poisson model should be used, which uses a Poisson count model and a logit model for

estimating the excess zeros (Lambert, 1992). In a frequentist setting, inference for GLMs is typically achieved by using iteratively reweighted least squares to obtain the maximum likelihood estimator (Dobson and Barnett, 2008), whereas in a Bayesian setting the Metropolis-Hastings algorithm can be used.

2.4 Geostatistical modelling

Geostatistical modelling is a way of describing spatial patterns and interpolating values for locations in space where data has not been observed. A geostatistical process is a realisation of the stochastic process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ where $\boldsymbol{s}$ are the locations where data could occur which vary continuously over the study region $D \subset \mathbb{R}^2$. However in practice, data are usually collected at a finite number of locations, $\boldsymbol{y} = \{y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_m)\}$.

A geostatistical process is defined to be Gaussian if the joint distribution of these observations is multivariate Gaussian. In this case the process is completley defined by its first two moments, $\mathbb{E}[y(\boldsymbol{s})] = \mu(\boldsymbol{s})$ and $C_y(\boldsymbol{s}, \boldsymbol{t}) = \text{cov}(y(\boldsymbol{s}), y(\boldsymbol{t}))$.

A geostatistical process is weakly stationary if

1. $\mathbb{E}[y(\boldsymbol{s})] = \mu_y(\boldsymbol{s}) = \mu_y$ for some finite constant μ which does not depend on $\boldsymbol{s}$.

2. $\text{Cov}[y(\boldsymbol{s}), y(\boldsymbol{t})] = C_y(\boldsymbol{s}, \boldsymbol{t}) = C_y(\boldsymbol{h})$, where $\boldsymbol{h} = \boldsymbol{s} - \boldsymbol{t}$.

Essentially this means that the mean is constant in space and the covariance function between two points depends only on the distance and direction between them, and not the locations themselves (Diggle and Ribeiro, 2007).

A weakly stationary geostatistical process is isotropic if the covariance function can be further simplified to:

$$C_y(\boldsymbol{h}) = C_y(||\boldsymbol{h}||), \tag{2.12}$$

where $h = ||\boldsymbol{h}||$ denotes the Euclidean distance of the spatial lag $\boldsymbol{h}$, i.e. the covariance function depends only on the distance between two points and not the direction (Diggle and Ribeiro, 2007).

2.4.1 Covariance functions

Suppose we have a Gaussian weakly stationary and isotropic process such as:

$$\boldsymbol{Y} \sim \mathrm{N}(\mu \mathbf{1}, \boldsymbol{\Sigma}(\boldsymbol{\lambda})). \tag{2.13}$$

The spatial autocorrelation in the data is estimated via the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\sigma^2, \tau^2, \phi)$. Here

- σ^2 is the partial sill and represents the spatial variation in the data.
- τ^2 is the nugget and represents the non-spatial variation in the data.
- ϕ is the range parameter which measures how quickly the covariance decays to zero.

Covariance functions can be used to account for the correlation between locations. The most commonly used covariance function in geostatistical modelling is the exponential covariance function (Diggle and Ribeiro, 2007), defined by

$$C_y(h) = \begin{cases} \sigma^2 \exp(-h/\phi), & h > 0 \\ \tau^2 + \sigma^2, & h = 0 \end{cases} \tag{2.14}$$

2.4.2 Kriging

Kriging was first proposed by Krige (1951) to predict spatial processes at new spatial locations, $\boldsymbol{s}_0$. The approach is based on deriving the Best Linear Unbiased Prediction (BLUP) for a new location given our current data $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_m))$. The BLUP can be obtained by minimising the mean square prediction error (MSPE):

$$\mathrm{MSPE} = \mathbb{E}[(y(\boldsymbol{s}_0) - \hat{y}(\boldsymbol{s}_0))^2]. \tag{2.15}$$

It can be shown that the MSPE is minimised at $\hat{y}(\boldsymbol{s}_0) = \mathbb{E}[(y(\boldsymbol{s}_0)|\boldsymbol{y}(\boldsymbol{s})]$. This allows for the application of the conditional distribution property of a multivariate Gaussian

distribution (see A.1). Applying this property gives the joint geostatistical process at the m data locations $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_m))$ and a new location $y(\boldsymbol{s}_0)$ as:

$$\begin{pmatrix} y(\boldsymbol{s}_0) \\ \boldsymbol{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y \\ \mu_y \mathbf{1} \end{pmatrix}, \begin{pmatrix} C_y(0, \boldsymbol{\lambda}) & \boldsymbol{C}_y(\boldsymbol{s}_0, \boldsymbol{\lambda})^\top \\ \boldsymbol{C}_y(\boldsymbol{s}_0, \boldsymbol{\lambda}) & \boldsymbol{\Sigma}(\boldsymbol{\lambda}) \end{pmatrix} \right), \tag{2.16}$$

where $C_y(0, \boldsymbol{\lambda}) = \mathrm{Var}[y(\boldsymbol{s}_0)]$, $\boldsymbol{C}_y(\boldsymbol{s}_0, \boldsymbol{\lambda}) = (C_y(||\boldsymbol{s}_1 - \boldsymbol{s}_0||, \boldsymbol{\lambda}), ..., C_y(||\boldsymbol{s}_n - \boldsymbol{s}_0||, \boldsymbol{\lambda})$ and $\boldsymbol{\lambda} = (\sigma^2, \tau^2, \phi)^\top$ are the parameters from the chosen covariance function $C_y()$. It then follows that:

$$\mathbb{E}[y(\boldsymbol{s}_0)|\boldsymbol{y}] = \hat{\mu}_y + \boldsymbol{C}_y(\boldsymbol{s}_0, \hat{\boldsymbol{\lambda}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}})^{-1}(\boldsymbol{y} - \hat{\mu}_y \mathbf{1}), \tag{2.17}$$

and,

$$\mathrm{Var}[y(\boldsymbol{s}_0)|\boldsymbol{y}] = C_y(0, \hat{\boldsymbol{\lambda}}) - \boldsymbol{C}_y(\boldsymbol{s}_0, \hat{\boldsymbol{\lambda}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}})^{-1} \boldsymbol{C}_y(\boldsymbol{s}_0, \hat{\boldsymbol{\lambda}}). \tag{2.18}$$

Equation 2.17 is called the Ordinary Kriging Predictor for an unknown, constant mean. The same approach can be used to find the Universal Kriging Predictor which can be used when the mean is non-constant and unknown.

2.5 Spatial modelling

Given that the direct modelling of the SIR can lead to unstable estimates of risk, a common approach is to extend the Poisson GLM (2.10) to account for spatial variation in the data. Given that these data are collected over space we would expect to see spatial correlation between areas which are spatially close together. Spatial correlation is induced into our modelling techniques via the neighbourhood matrix $\boldsymbol{W}$, which is an $(n \times n)$ normally binary matrix, where $w_{ij} = 1$ if two areas are defined to be neighbours and $w_{ij} = 0$ if not. Given that an area cannot be neighbours with itself, $w_{ii} = 0$ for all i. There are many ways to define if areas i and j are neighbours, for example if they share a common border, if the centroids of the areas are within a fixed distance d of each other or if area i is one of the k closest areas to area j in terms of distance. In this thesis we define two areas to be neighbours if they share a

common border.

The most common statistic to measure spatial correlation is Moran's I (Moran, 1950), which is defined as follows:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{2.19}$$

Moran's I measures the strength of the linear spatial association in the areal data, suitably weighted for their proximities. Potentially $-1 < I < 1$, where if $I = -1$ we have perfect dispersion, if $I = 0$ we have a random arrangement and if $I = 1$ we have perfect positive correlation. In reality we would expect a positive value of Moran's I, since areas close together are generally more likely to have similar values. A permutation test is carried out to test the null hypothesis of no spatial correlation, where the observed Moran's I statistic is compared to Moran's I statistics based on K different random permutations of the data (which should give K values of Moran's I under independence). The estimated two-sided p-value for the test is

$$\frac{2}{K+1} \sum_{k=1}^{K} I(I_k > |I_{obs}|). \tag{2.20}$$

Here I_{obs} is the observed Moran's I test statistic and $I_1, ..., I_k$ are the Moran's I statistics based on the K different random permutations of our data.

If the data contains significant spatial correlation, a popular way to model these data is through a hierarchical Bayesian model with inference based on McMC simulation. Given the convenient structure of hierarchical models, spatial correlation can be incorporated easily by extending the simple Poisson generalised linear model in Section 2.3 to a generalised linear mixed model with a set of spatially varying random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ by

$$\begin{aligned} Y_i &\sim \text{Poisson}(e_i\theta_i), \\ \ln(\theta_i) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \phi_i, \end{aligned} \tag{2.21}$$

where $\boldsymbol{x}_i = (1, x_{i2}, ..., x_{ip})$ is a $p \times 1$ vector of known covariates, including an intercept term, with regression parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. Commonly the spatial random effects are modelled via a conditional autoregressive (CAR) prior, which can be defined by set of univariate full conditional distributions of the form $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. Many different CAR models have been proposed and four of the most popular are detailed in the following sections.

2.5.1 Intrinsic CAR

The first CAR model prior was developed by Besag et al. (1991) and is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right). \tag{2.22}$$

Given that w_{ij} is only equal to 1 if two areas are defined to be neighbours, then the conditional expectation of data point i is the mean of the data points in neighbouring areas, and so each area is modelled as being similar to its neighbours. The conditional variance also decreases as the number of neighbours increases, i.e. the more neighbours an area has the lower the variance. This is natural since the more neighbours an area has the more information there is about that area. Although this seems a natural way to model these data, it only models strong correlation and does not contain a correlation parameter which would allow for the strength of spatial correlation to be estimated from the data. This prior is also undetermined for singleton areas.

2.5.2 Convolution CAR

The convolution CAR was also delevoped by Besag et al. (1991) and is given by

$$\begin{aligned}\phi_i &= \phi_i^{(1)} + \phi_i^{(2)} \\ \phi_i^{(1)}|\boldsymbol{\phi}_{-i}^{(1)} &\sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_j^{(1)}}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau_1^2}{\sum_{j=1}^{n} w_{ij}}\right) \\ \phi_i^{(2)} &\sim N\left(0, \tau_2^2\right).\end{aligned} \tag{2.23}$$

Here $\boldsymbol{\phi}$ is now a linear combination , with $\boldsymbol{\phi^{(1)}}$ assigned the intrinsic CAR prior and the second set of random effects, $\boldsymbol{\phi^{(2)}}$, being independent and identically distributed with mean zero and constant variance. This then overcomes the issue of the intrinsic CAR prior inducing too much spatial smoothness by allowing this level to be determined by the ratio of the two variances τ_1^2/τ_2^2. However, the disadvantage of this model is that each data point is now represented by two random effects and only the sum, $\phi_i^{(1)} + \phi_i^{(2)}$, is identifiable.

2.5.3 Proper CAR

The model proposed by Stern and Cressie (2000) uses a single set of random effects but allows for the level of spatial correlation to be estimated by introducing a correlation parameter ρ. The full conditional distributions are given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right). \tag{2.24}$$

Here ρ controls the level of spatial correlation in the data with $\rho = 0$ corresponding to independence in space and $\rho = 1$ corresponding to strong spatial dependence. The conditional variance is the same as the intrinsic CAR model. This model is essentially the spatial equivalent of an AR(1) process in time series. The main issue with this model is that when ρ is close to zero (indicating near independence in space), the conditional variance is inversely proportional to the number of neighbours an area has. However, ideally, in the absence of spatial correlation, the conditional variance would

not decrease as the number of neighbours increases since having more neighbours does not decrease the uncertainty about the value of an areas random effect.

2.5.4 Leroux CAR

This issue was addressed in the model proposed by Leroux et al. (2000) which is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}\right), \tag{2.25}$$

again ρ controls the level of spatial correlation in the data. However now if $\rho = 0$, the conditional variance simplifies to a constant, τ^2, and therefore the variance will remain constant regardless of the number of neighbours an area has.

2.6 Spatio-temporal modelling

Now suppose that for each area in region $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$, the response is observed over a period of $t = 1, .., T$ years, resulting in the response vector $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iT})$ for each area $\mathcal{A}_i$. Data of this type are very common and lead not only to correlation in space but also correlation in time.

The notion of separability is important in spatio-temporal statistics. If the simplifying assumption of separability is made, the spatio-temporal covariance structure can be separated into the product of two functions, one which depends only on space and one which depends only on time. The following section describes some of the spatio-temporal models proposed, of which some make this simplifying assumption. It should also be noted that these models are described without covariates, however the addition of these is trivial.

2.6.1 Bernardinelli model

One of the first papers on space-time disease mapping was by Bernardinelli et al. (1995), who proposed a hierarchical Bayesian model for the analysis of risk for a given disease over space and time. The model proposed allows for analysis of data collected over space and time by introducing a time effect along with a spatial effect. The model proposed for θ_{it} (disease risk), where i $(i = 1, ...n)$ is the area and t $(t = 1, ..., T)$ is the time point, is as follows:

$$\log(\theta_{it}) = \mu + \phi_i + (\beta + \delta_i)t. \tag{2.26}$$

Here μ is the overall intercept term and β is the overall slope parameter for the linear time trend. The intercept can vary over space by the introduction of the random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, as can the linear slope parameters via $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$. Essentially, this model allows for a separate intercept and linear time trend for each spatial area and hence a non-separable space-time structure.

Two separate prior distributions for ϕ_i and δ_i are proposed. The first is a normal distribution with mean 0 and variance σ^2 used to model spatially unstructured variation (i.e. independence). The second is an Intrinsic CAR prior used to model spatially structured variation. This model also allows for correlation between the slope and intercept by using an additional level in the hierarchical model. However it should be noted that it may be inappropriate to restrict the time structure to be linear, especially over long periods of time.

2.6.2 McNab and Dean model

An extension of this model was presented by MacNab and Dean (2002), who proposed the use of B-spline trends over the temporal component. The model is of the form:

$$\log(\theta_{it}) = \mu + \phi_i + \alpha(t) + \beta_i(t), \tag{2.27}$$

where μ is the overall mean, ϕ_i is a random effect for area i, $\alpha(t)$ is an overall time trend for all areas while $\beta_i(t)$ is an area specific deviation from this overall trend.

As usual, a CAR prior is used for the spatial random effects and two specifications are considered for modelling the temporal effects. Firstly, the overall trend, $\boldsymbol{\alpha}$, is modelled using a cubic B-spline and the area specific trends are assumed to be linear. The second proposed form uses cubic B-splines to model both components.

The models proposed here provide a flexible approach to modelling data with complex spatio-temporal correlation structures. The use of both overall and area specific trends allows for the estimation of the overall temporal structure observed across the spatial region and the temporal structure of each individual area separately, to identify areas which have a substantially different temporal trend than the overall mean. Modelling the $\beta_i(t)'s$ linearly, where appropriate, simplifies the model considerably and allows for the random effects to be expressed linearly as $\phi_i + \beta_i t$, which has a simple interpretation with ϕ_i corresponding to the random effect over space and β_i being the temporal linear trend for area i on top of the overall temporal trend (reducing to the Bernardinelli model).

However, the increased flexibilty of using B-spline smoothers for both the overall and area specific trends results in a large increase in the number of parameters that need to be estimated. Given that there are several possible alternatives for modelling the temporal random effects, choices need to be made about which of the two temporal structures to use. Also, for effects modelled using B-splines, the degree of polynomial to be used must be chosen as well as how many and where to place the interior knots.

2.6.3 Knorr-Held model

The paper by Knorr-Held (2000) proposed a space-time hierarchical Bayesian model with main effects and a space×time interaction. Rather than assume that the number of disease cases or deaths Y_{it}, for area i and time point t, follows a Poisson distribution, Y_{it} is assumed to have a binomial distribution with parameters n_{it} and π_{it} being the number of persons at risk and the underlying binomial probability both for area i and time point t respectively. The model proposed is as follows:

$$\text{logit}(\pi_{it}) = \mu + \phi_i + \alpha_t + \gamma_t + \theta_i + \delta_{it}, \tag{2.28}$$

where μ is the overall intercept, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_T)$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_T)$ are temporal effects and $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ and $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are spatial effects. Main effects $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ have no temporal and spatial structure *a priori*, whereas blocks $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ do have temporal and spatial structure *a priori*, this model can therefore be thought of as the spatio-temporal equivalent of the Convolution model described in Section 2.5.2. The same model was proposed by Knorr-Held and Besag (1998) except for the introduction of the interaction term $\boldsymbol{\delta} = (\delta_{11}, ..., \delta_{nT})$. The motivation for this more complex model was to try and improve upon Bernardinelli et al. (1995) with a less restrictive temporal trend and Knorr-Held and Besag (1998) with the additional allowance of a space×time interaction term to allow for cases where variation in disease cannot be modelled using separate space and time main effects, but an interaction between these is necessary. The interaction parameters δ_{it} can follow one of four different formulations, each corresponding to a different degree of prior dependence. These are: (i) independent over space and time; (ii) independent over space but dependent over time; (iii) independent over time but dependent over space; and finally (iv) dependent over space and time. If all $\delta_{it} = 0$, this term can be removed and a separable model results.

This model formulation is beneficial because it is less restrictive than that proposed by Bernardinelli et al. (1995) and also introduces a space×time interaction which was not possible in the previous model proposed by Knorr-Held and Besag (1998). It also allows for several space×time structures to be modelled by introducing four types of prior distribution for $\boldsymbol{\delta}$. Modifications can also be made to the general specification if required, for example it may not be necessary to allow for both structured and unstructured variation in space and time, which are computed as Kronecker products of the spatial and temporal precision matrices. The interaction term is also included in a way that allows the model to be simplified to just the main effects if $\boldsymbol{\delta}$ turns out to be negligible.

However, the general model formulation contains a large number of parameters, four main effects plus one interaction, and therefore five variance parameters and there may not be enough information in the data to reliably estimate them all.

2.6.4 Ugarte model

Ugarte et al. (2012) propose a similar model to that proposed by Knorr-Held (2000) which is as follows:

$$\log(\theta_{it}) = \mu + \alpha_t + \phi_i + \delta_{it}, \tag{2.29}$$

where μ is the overall intercept, α_t are temporal effects, ϕ_i are spatial effects and δ_{it} are space-time interactions. CAR type distributions for the spatial, temporal and spatio-temporal interaction terms are assumed, namely a first order random walk for α_t, a Leroux CAR prior for ϕ_i and a normal prior for δ_{it} with mean 0 and a precision matrix which is the Kronecker product of the precision matrices for the other two effects. This formulation is more parsimonious than that proposed by Knorr-Held (2000) as there are 2 fewer variance parameters to be estimated and 2 fewer sets of random effects. Separability can also be tested by checking if the variance component for $\boldsymbol{\delta}$ is estimated to be 0 and if so, the interaction term can be dropped.

However, unlike the spatial effect, the temporal effect does not contain a correlation parameter to allow for the strength of the temporal correlation to be estimated and so this formulation is only suitable for modelling strong temporal correlation.

2.6.5 Rushworth model

The model proposed by Rushworth et al. (2014) is a less highly parametrised extension of Ugarte et al. (2012) and is of the form:

$$\log(\theta_{it}) = \phi_{it}, \tag{2.30}$$

where ϕ_{it} are spatio-temporal random effects. Temporal correlation is induced by allowing $\boldsymbol{\phi}_t = (\phi_{1t}, ..., \phi_{Nt})$, to depend on $\boldsymbol{\phi}_{t-1}$. The set of random effects for time point 1, $\boldsymbol{\phi}_1$, is specified using a Leroux CAR prior since $\boldsymbol{\phi}_0$ does not exist and the conditional specification for all other random effects is given by

$$\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} \sim N(\alpha\boldsymbol{\phi}_{t-1}, \tau^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}) \quad t = 2, ..., T, \tag{2.31}$$

where the precision matrix $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$, where $\boldsymbol{I}$ is the $N \times N$ identity matrix and $\mathbf{1}$ is an $N \times 1$ vector of 1s. Unlike the model proposed by Ugarte et al. (2012), here α controls the level of temporal correlation which gives this formulation the increased flexibility of modelling varying degrees of temporal correlation. Another advantage is that only one variance parameter has to be estimated as there is only one set of random effects used to capture residual spatio-temporal correlation. In all other formulations, a variance parameter would have to be estimated for each set of random effects.

However this particular specification can only be used for non-separable space-time data given that the space, time and space×time effect are all represented by one random effect which cannot simply be dropped if it turns out to be unnecessary. There are also no space and time random effects to capture overall trends in the data.

2.7 Multivariate spatial models

So far the models introduced use data collected on one response variable of interest. However, it may be of interest to model several response variables simultaneously using multivariate techniques. Now suppose that for each area in the study region, $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$, more than one response is observed for each area, say $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iD})$, where $d = 1, ..., D$ denotes the different responses. Many multivariate models have been proposed in a purely spatial setting using multivariate CAR (MCAR) spatial models, some of which are detailed in the following sections.

2.7.1 Kim model

Firstly, CAR models were extended by Kim et al. (2001) to the bivariate setting, by proposing the following model with a two-fold CAR prior for the spatial effects

$$\log(\theta_{id}) = \psi_d + \phi_{id} + \epsilon_{id}, \tag{2.32}$$
$$\boldsymbol{\phi} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right),$$

where θ_{id} is the mortality rate for area $i = (1, ..., n)$ and outcome $d = (1, 2)$, ψ_d is a disease specific intercept, ϕ_{id} is the spatial effect of the i^{th} region for the d^{th} outcome, and ϵ_{id} is the extra variation effect where $\epsilon_{id} \sim N(0, \sigma_d^2)$ with σ_d^2 assumed known. The spatial effects $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$, where $\boldsymbol{\phi}_1 = (\phi_{11}, ..., \phi_{n1})$ and $\boldsymbol{\phi}_2 = (\phi_{12}, ..., \phi_{n2})$, follow a multivariate normal with mean $\mathbf{0}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}^{-1}$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \frac{1}{\delta_1}(\boldsymbol{D} - \rho_1 \boldsymbol{W}) & -\frac{1}{\sqrt{\delta_1 \delta_2}}(\rho_0 \boldsymbol{I} + \rho_3 \boldsymbol{W}) \\ -\frac{1}{\sqrt{\delta_1 \delta_2}}(\rho_0 \boldsymbol{I} + \rho_3 \boldsymbol{W}) & \frac{1}{\delta_2}(\boldsymbol{D} - \rho_2 \boldsymbol{W}) \end{pmatrix}. \tag{2.33}$$

Here $\boldsymbol{W}$ is the usual neighbourhood matrix and $\boldsymbol{D}$ is an $(n \times n)$ diagonal matrix defined by $\text{diag}(2d_1 + 1, ..., 2d_n + 1)$, where $d_i = \sum_{j \ i} w_{ij}$, the number of neighbours of region i. However this model is designed for the bivariate case and is not easily generalised to higher dimensions.

2.7.2 Gelfand model

A more general formulation of MCAR models was proposed by Gelfand and Vounatsou (2003), who proposed a number of models, one of the simplest given by

$$\begin{aligned} \log(\theta_{id}) &= \phi_{id}, \quad i = 1, ..., n, d = 1, ..., D, \\ \boldsymbol{\phi} &\sim N\left(\mathbf{0}, \left[\boldsymbol{Q}(\boldsymbol{W}, \rho) \otimes \boldsymbol{\Sigma}^{-1}\right]^{-1}\right), \end{aligned} \tag{2.34}$$

where $\boldsymbol{\Sigma}$ is the between disease covariance matrix and the spatial correlation is induced via the precision matrix

$$\boldsymbol{Q}(\boldsymbol{W}, \rho) = [\text{diag}(\boldsymbol{W}\mathbf{1}) - \rho \boldsymbol{W}]. \tag{2.35}$$

This MCAR model can be thought of as the multivariate equivalent of the proper CAR detailed in Section 2.5.3. For a more comprehensive summary of the MCAR literature refer to MacNab (2016).

2.8 Multivariate spatio-temporal models

A natural extension of the MCAR models (as described in Section 2.7) is to allow for multivariate data collected not only in space but also over time. Now assume that for every area in region $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$, a response Y_{itd} is observed for area $i = 1, ..., n$, time period $t = 1, ..., T$ and outcome $d = 1, ..., D$. Modelling techniques for data of this kind will allow for information to be shared over space, time and outcome, and should ensure that spatial, temporal and outcome correlation in the data are accounted for. So far, the literature extending MCAR models to allow for data collected over time is very limited. The following section contains details on some of the multivariate spatio-temporal modelling techniques which have been proposed.

2.8.1 Tzala and Best model

Of the few models proposed, some adopt Bayesian latent variable modelling. The general idea behind this is to use correlation or covariances among a set of observed variables and describe them in terms of a smaller set of latent variables. This approach was adopted by Richardson et al. (2006) to model two diseases only. This framework was extended to allow for more than two outcomes by Tzala and Best (2008), who used factor analytic models applied to six diet-related cancers in Greece. The model proposed is as follows:

$$\log(\theta_{itd}) = \mu_{itd}, \tag{2.36}$$

where the latent common factor is introduced as part of the model for μ_{itd}. The paper gives three main formulations for this part of the model, the simplest of which is a representation of a simple factor analysis model given by

$$\mu_{itd} = \lambda_d f_{it}. \tag{2.37}$$

Here f_{it} is the latent common factor and λ_d represents the factor loading for disease d. The second model formulation proposed in Tzala and Best (2008) extends this model

by allowing for disease specific spatial and temporal trends. In the third proposed model in this paper, this formulation is extended by separating the common factor, λ_d, into two individual components, one each for spatial and temporal structure separately.

2.8.2 Quick model

An alternative to latent variable modelling is to extend the MCAR literature to allow for data collected over space and time. A non-separable multivariate spatio-temporal Bayesian model was proposed by Quick et al. (2017a). However, rather than model the counts directly using a Poisson likelihood, they model the log of the disease rates as Gaussian to avoid the computational burden that is associated with a Poisson model. The proposed model, which follows on from Gelfand and Vounatsou (2003) described in Section 2.7.2, is termed a multivariate space-time CAR (MSTCAR) and is of the form:

$$\ln(\theta_{itd}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_d + \mathcal{Z}_{itd} + \phi_{itd}, \tag{2.38}$$

$$\boldsymbol{Z} \sim N\left(\mathbf{0}, [\boldsymbol{Q}(\boldsymbol{W}) \otimes \boldsymbol{\Sigma}_\eta^{-1}]^{-1}\right), \tag{2.39}$$

where $\boldsymbol{Z}$ is a vector of random effects which account for spatio-temporal and between outcome dependence, and $\phi_{itd} \sim \mathrm{N}(0, \tau_d^2)$. Spatial correlation is induced via the precision matrix

$$\boldsymbol{Q}(\boldsymbol{W}) = [diag(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}], \tag{2.40}$$

which is the multivariate equivalent to the Intrinsic CAR model described in Section 2.5.1. Here $\boldsymbol{\Sigma}_\eta$ is the LDU decomposition (decomposition of the form $A = LDU$ where D is a diagonal matrix and L and U are lower and upper diagonal matrices respectively) comprising of temporal correlation and between outcome correlation and allows for group-specific temporal correlation parameters and temporally varying covariance matrices. Although using the normal approximation allows for computa-

tional efficiency, it may not be appropriate if modelling data containing low rates of incidence. This model was extended in Quick et al. (2017b) to a generalised linear model setting to analyse age-specific stoke mortality data with a Poisson likelihood.

2.9 Continuous inference on aggregated data

So far, all of the models described make inference on the spatial distribution of disease risk for aggregated data relating to irregularly shaped non-overlapping areal units. Data of this type are very common in disease mapping due to confidentiality issues and so traditionally, models for discrete spatial variation, such as the CAR models described earlier, are adopted to directly model the aggregated counts. Although these models can be extremely useful to health authorities, one shortcoming of traditional areal unit modelling techniques is that the areal units are themselves artificial units of spatial recording and can influence the spatial pattern observed in the data. That is, if the areal units changed then so would the results. This is known as the Modifiable Areal Unit Problem (MAUP) and is a well known issue with discrete spatial modelling (Heywood et al., 1998). Several studies of the MAUP have been conducted (Fotheringham and Wong, 1991; Jelinski and Wu, 1996) which give evidence of unreliability of analysis undertaken with data from areal units. Another common problem in areal unit data of this type is that often there are changes to boundaries that occur during the time period for which data are available. For example, in 2014 The Scottish Government released a redrawn version of the intermediate geography boundaries, and there are several data sets publicly available for which the time period overlaps this boundary change. Using data from before and after this change would lead to incomparable inference due to spatial misalignment in the data, and needs to be dealt with. One way to overcome this issue would be to undertake inference on a common latent spatial grid scale, and use a data augmentation approach to estimate the unknown grid level counts on this scale. These grids can be made arbitrarily small which leads to approximate continuous inference. From a Bayesian perspective several approaches for this have been proposed.

The aim of Li et al. (2012a) was to make inference on the spatial distribution of

syphilis risk in North Carolina using log-Gaussian Cox processes (LGCPs) to construct spatially continuous maps of disease risk with the process being modelled on a fine grid rather than the original census tracts. The spatial risk surface is modelled as:

$$\log(\theta(s)) = \mu + X(s)\beta + U(s), \tag{2.41}$$
$$\text{Cov}[U(s), U(s+h)] = \sigma^2 \text{Matérn}(|h|/\phi, \nu),$$

where $\theta(s)$ is the risk surface over space, $X(s)$ is a vector of covariates measured at each location in space and $U(s)$ are spatial random effects. Here ϕ, ν are the range and roughness parameters respectively and h is the distance between two points. The continuous risk surface $U(s)$ is approximated by a piecewise constant surface evaluated on a regular lattice of squared grid cells. A data augmentation step allocates the observed disease counts, Y_r into counts per intersection of cell and areal units and these are modelled as Poisson. A similar approach to these models was also taken by Diggle et al. (2013).

The model proposed by Taylor et al. (2017) extends these models by adopting spatio-temporal log-Gaussian Cox processes proposed in Taylor et al. (2015) and applying these to aggregated areal unit data. The model proposed allows the risk surface to vary over space and time as:

$$\log(\theta(s,t)) = X(s,t)\beta + U(s,t), \tag{2.42}$$

where $\theta(s,t)$ is the risk surface over space and time, $X(s,t)$ is a vector of covariates measured at each location in space and time and $U(s,t)$ are spatio-temporal random effects. This paper also extends the data-augmentation methods developed in Li et al. (2012a) and Taylor et al. (2015) to allow for continuous inference on spatio-temporal areal unit data with overlapping and uncertain boundaries. Again inference is made on a fine computational grid and there are many ways that the continuous spatio-temporal surface could be modelled.

For example Rostami et al. (2017) proposed a log-Gaussian Cox process to model spatio-temporal variation of substance abuse mortality in Iran. Here a seperable spatio-temporal risk surface was implemented of the form:

$$\text{Cov}[U(\boldsymbol{s},t),U(\boldsymbol{s}',t')] = \sigma^2\text{Matérn}(||\boldsymbol{s}-\boldsymbol{s}'||)\rho^{|t-t'|}, \tag{2.43}$$

where $||\boldsymbol{s}-\boldsymbol{s}'||$ is the Euclidean distance between $\boldsymbol{s}$ and $\boldsymbol{s}'$ and ρ denotes the temporal correlation.

Chapter 3

A single disease spatio-temporal model to estimate changes in health inequalities in coronary heart disease across Scotland.

3.1 Introduction

The motivation for the work in this thesis is to estimate health inequalities across Scotland. Although there have been many previous studies looking at health inequalities in Scotland, many of them focus on comparisons either between large areas within Scotland or between Scotland and other western European countries. Therefore, in this chapter we propose a spatio-temporal model for quantifying health inequalities in one disease in Scotland at a small area level using disease mapping techniques. The disease we will focus on in this chapter is coronary heart disease, which is one of the biggest killers in Scotland (Scotpho, 2016). The main focus is answering several questions of interest:

1. Are there health inequalities in risk of coronary heart disease between Scotland's 14 regional health boards and how are these changing over time?

2. How are health inequalities changing over time in IGs in Scotland for coronary heart disease risk?

3. What impact do the covariates have on disease risk?

We will present the results from our study and answer these questions of interest in Section 3.5. However, first the data are presented in Section 3.2, while our proposed model with health board specific auto-regressive temporal effects and a baseline CAR prior for the spatial effects is presented in Section 3.3. Finally, Section 3.6 provides a discussion on the conclusions drawn from this study and how it will be developed in Chapter 4.

3.2 Data

3.2.1 Study region

As described in Chapter 1, the study region is Scotland which is split into $n = 1235$ intermediate geographies (see Figure 1.3) and $H = 14$ health boards (see Figure 1.4).

3.2.2 Disease data

The disease data are yearly counts of the numbers of hospital admissions for coronary heart disease for the years 2003 to 2012 in each IG. For each year and intermediate geography (IG) we have the number of admissions to non-psychiatric/non-obstetric hospitals in Scotland with a main diagnosis of coronary heart disease which is defined using the International Classification of Diseases Volume 10 (ICD10) codes (I20:I25).

The 10-year time period was chosen for a few reasons. Firstly, the data for these years are freely available from the Scottish Statistics website, which is available online at `http://statistics.gov.scot/`. Secondly, during this period The Smoking Health and Social Care (Scotland) Act 2005 banned smoking in any enclosed public space in Scotland from 26 March 2006 which is of interest since smoking has been proven to increase the likelihood of an individual developing heart disease (U.S Department of Health and Human Services, 2014). Finally, in 2007 The Scottish Government set up a Ministerial Task Force for Health Inequalities. Given that both

of these initiatives fall within or time period, we are interested in assessing if either of them had a positive effect on reducing the health inequalities in coronary heart disease in Scotland.

In order to adjust for age and sex differences in the populations in each IG, the expected numbers of hospital admissions were calculated separately for each disease using indirect standardisation based on age and sex adjusted rates for the whole of Scotland. Given one of the goals of this analysis is to investigate temporal trends in disease risk, the rates for the year 2006/07 were used to calculate the expected values for all years. This will ensure that any changes in risk of coronary heart disease are detected by the model rather than being incorporated into the expected values. There is no reason to believe that this choice will have an impact on the results since the disease studied here is chronic and therefore the risk for a population is unlikely to change dramatically year to year. The year 2006/07 was chosen since it lies in the middle of our time period. Letting i denote IG ($i = 1, ..., 1235$) and t denote year since 2003 ($t = 1, ..., 10$), the simplest measure of disease risk is the standardised incidence ratio (SIR), $\hat{\theta}_{it} = Y_{it}/e_{it}$, where Y_{it} is the observed number of hospitalisations and e_{it} is the expected number of hospitalisations. Values of SIR greater than 1 represent elevated levels of disease risk, and values less than 1 correspond to decreased levels of disease risk, for example, an SIR of 1.2 corresponds to a 20% increase in risk of coronary heart disease. Figure 3.1 shows the SIR for coronary heart disease admissions in IGs in Scotland from 2003 to 2012 firstly by year (panel(a)) and then by health board (panel(b)). From the top plot, a decreasing trend in coronary heart disease risk can be seen over the time period. In 2003 the median SIR is 1.177, whereas in 2012 this reduces to 0.765. There also seems to be a narrowing in the width of the boxplots suggesting that the overall inequality in coronary heart disease risk may be decreasing over time. When split by health board some variation in SIR can be seen, with some HBs showing more spread in SIR than others and some differences in the median level. In particular, the median SIR for Shetland (Z) is 0.666, and in contrast the median SIR for Lanarkshire (L) is 1.097.

Figure 3.2 shows boxplots of the SIR in IGs for each HB at each time point. From this it can clearly be seen that the risk of coronary heart disease is not constant

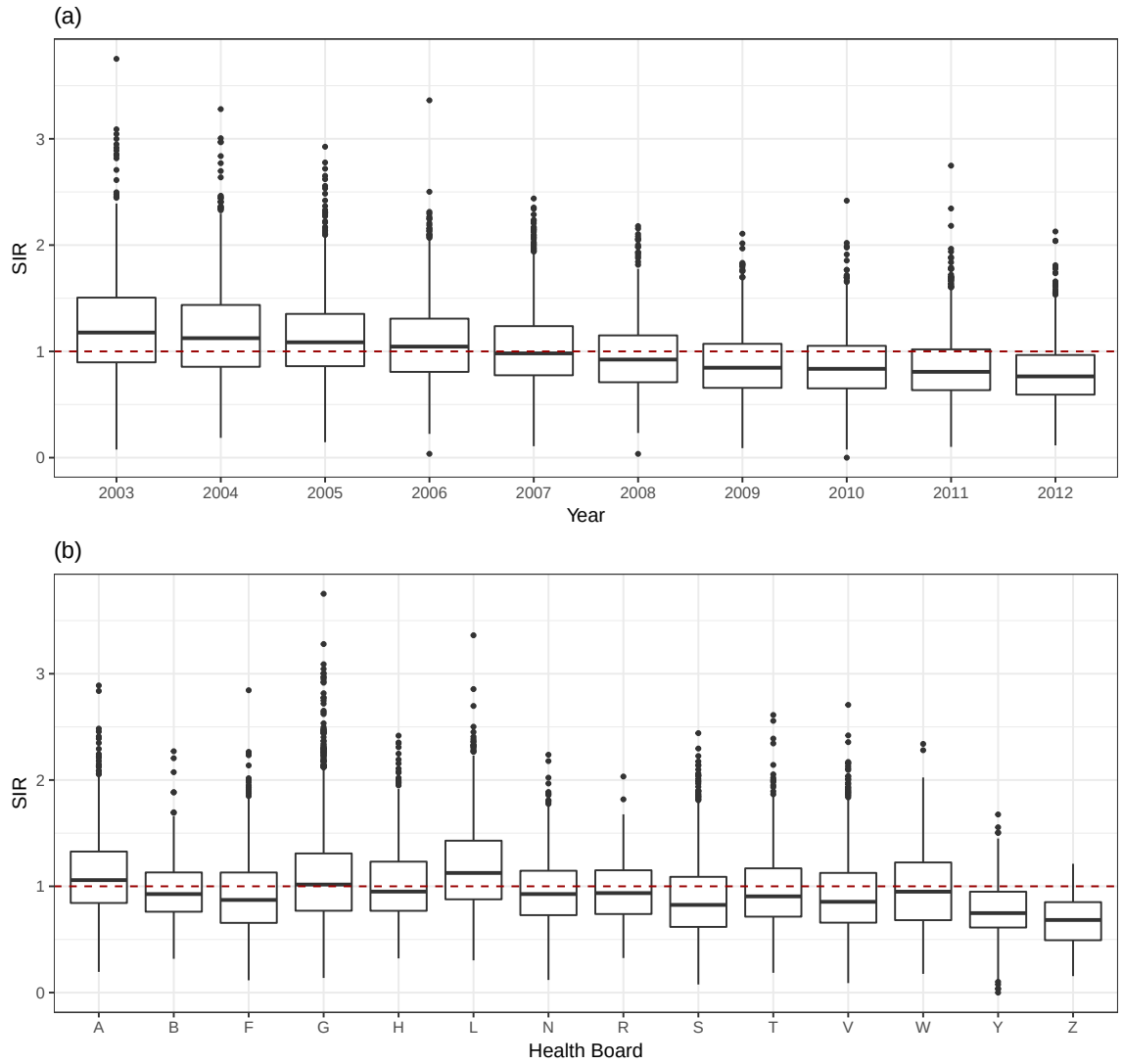

Figure 3.1: (Top panel (a)) Boxplots of the standardised incidence ratio (SIR) for coronary heart disease admissions for IGs in Scotland from 2003 to 2012 by year. (Bottom panel (b)) Boxplots of SIR for coronary heart disease admissions for IGs in Scotland from 2003 to 2012 by health board. Red dashed line indicates a risk of 1.

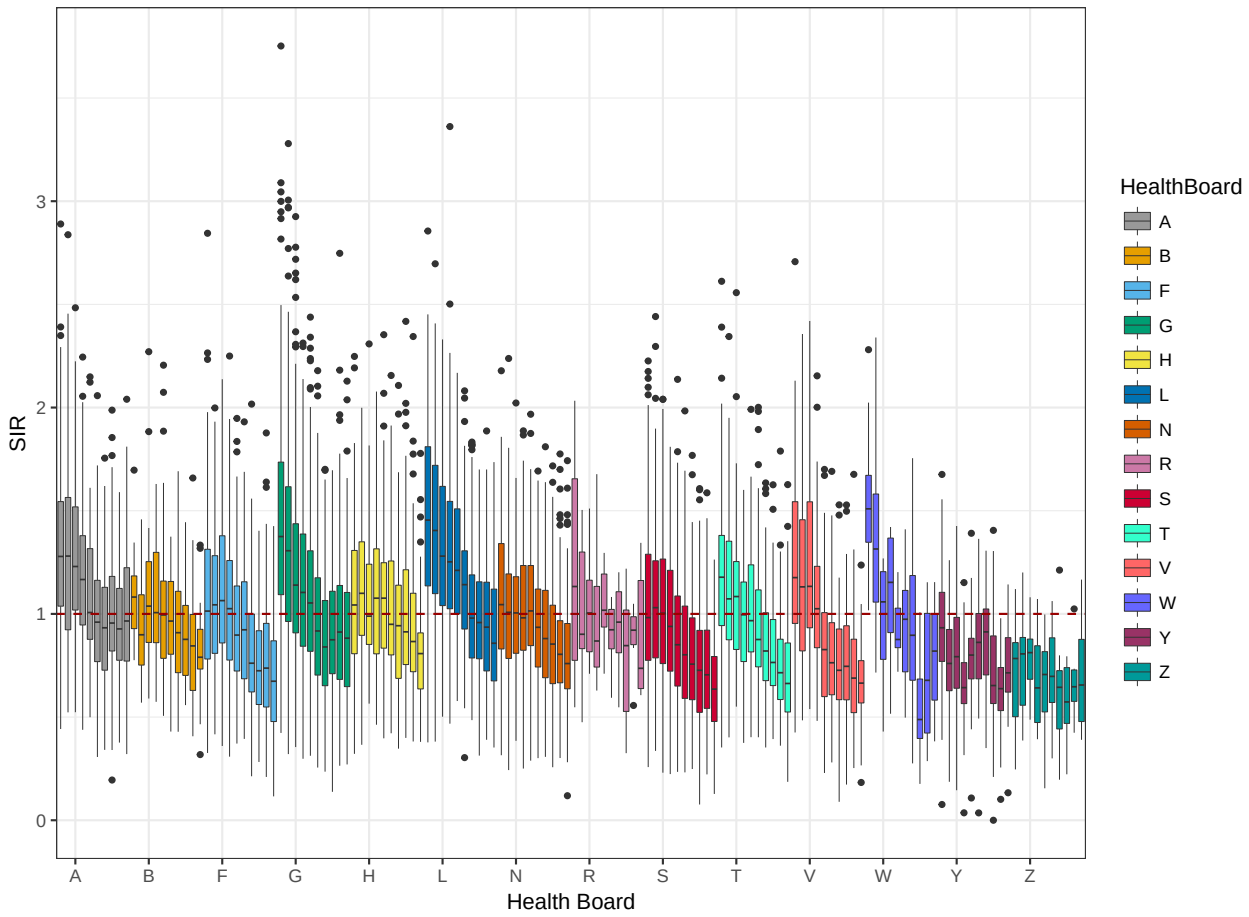

Figure 3.2: Boxplots of SIR for IGs in each health board at each year (2003-2012).

within a HB over the time period, for most HBs we see a decreasing trend over time. However, the extent of the reduction in SIR differs between the HBs, with some HBs showing a much stronger decrease than others. It can also be seen that the smaller health boards such as Orkney (R) and Western Isles (W) show much larger changes in the median level of SIR over time. This is due to these HBs having a very small number of IGs (see Table 1.1) and so the number of admissions for these HBs are small and therefore the data are more likely to show higher variation. From these plots, there certainly seem to be differences in the risk of SIR between the health boards and over time within health boards.

In order to assess the presence of spatial variation in the data, and how this has changed over the time period, Figure 3.3 shows the SIRs across IGs in Scotland in 2003 and 2012. Firstly, the map of SIR in 2003 appears to have more areas with high values of SIR than the map of 2012. This again suggests that risks of coronary heart disease are going down from 2003 to 2012. Given that it is difficult to see any pattern in the center of Scotland due to the high density of IGs, separate maps for

the health boards Greater Glasgow and Clyde (G), Lothian(S) and Lanarkshire (L) are shown in Figure 3.4. The decrease in overall risk is even more apparent from these plots. The numbers of areas in 2012 with high risk has reduced significantly compared to 2003, particularly in the west. This suggests that the decrease in risk of coronary heart disease may be more rapid for areas in this part of Scotland compared to the rest of Scotland. In both maps there are also far more areas with high risks in the west of central Scotland compared to the east. Most of the areas in the east of Glasgow city have SIRs of greater than one which is not surprising given that many areas there are highly deprived. In general, Edinburgh seems to have fewer areas with SIRs of greater than 1 compared to Glasgow. This stark difference in coronary heart disease incidence between Scotland's two major cities is a clear indication of the issue of health inequality across Scotland.

3.2.3 Covariate data

Potential covariates were identified to help describe the spatial variation in disease risk across Scotland. Firstly, the percentage of 16-64 year olds claiming job seekers allowance (JSA), which is an unemployment benefit that can be claimed while looking for work in the United Kingdom, is used as a proxy measure of deprivation since it is well known that higher levels of socio-economic deprivation is linked to increased disease risk (Audit Scotland, 2012). Given there is also evidence that a person's ethnicity can have an impact on the risk of certain diseases (The Scottish Government, 2012), the percentage of the population of Asian ethnicity and the percentage of the population of Black ethnicity were also included as potential explanatory variables. Both of these covariates are highly skewed to the right with lots of near zero values, and so a log transformation was applied. Finally, an urban/rural factor was included using the Scottish Government's urban rural 2-fold classification which can be found at `http://www.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification`. This was chosen as an indication of access to hospitals as perhaps those who live in rural areas are less likely to be admitted to hospital if they live in remote areas where hospitals are difficult to reach.

Figure 3.5 shows plots of the four potential covariates and SIR. The top left figure

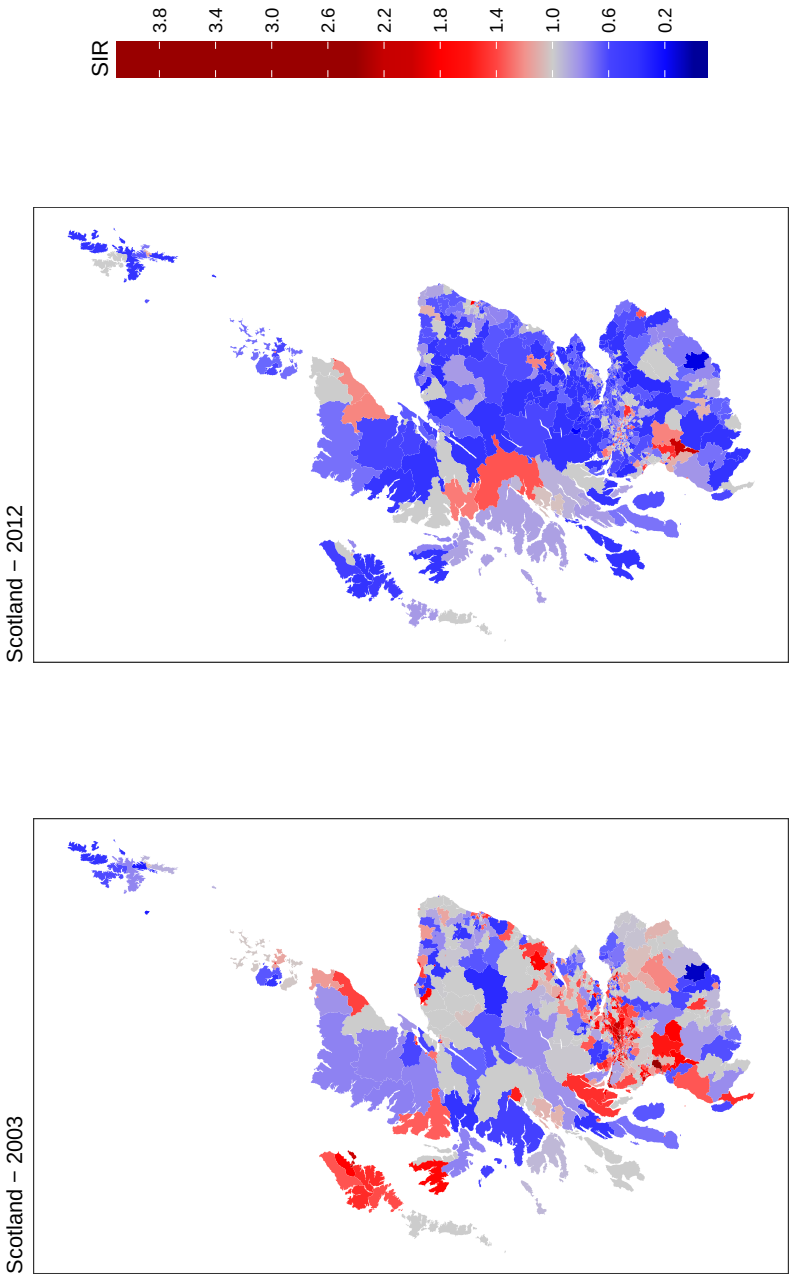

Figure 3.3: SIR for coronary heart disease for each IG in Scotland in 2003 and 2012.

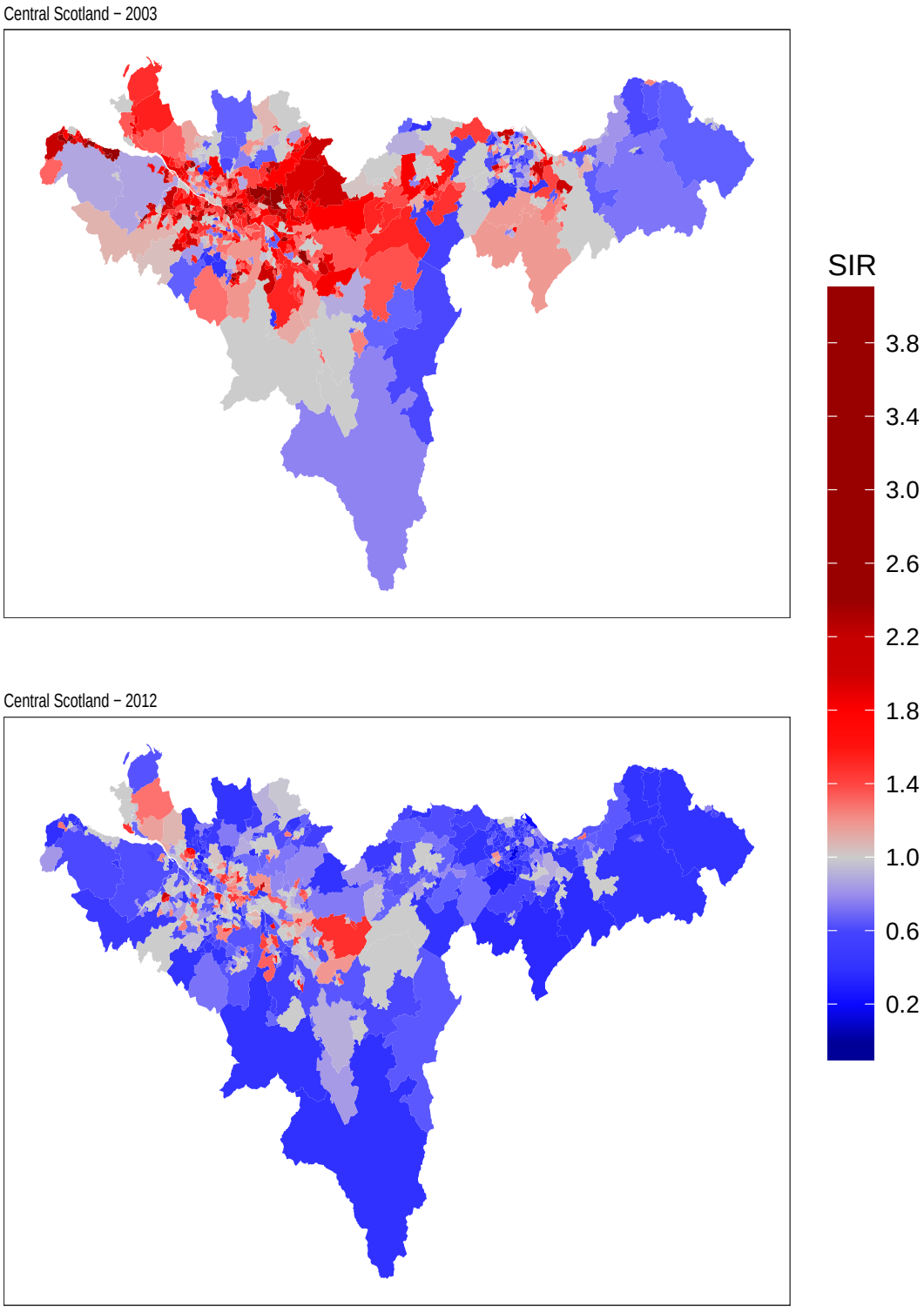

Figure 3.4: SIR for coronary heart disease for each IG in health boards Greater Glasgow and Clyde (G), Lothian (S) and Lanarkshire (L) in 2003 and 2012

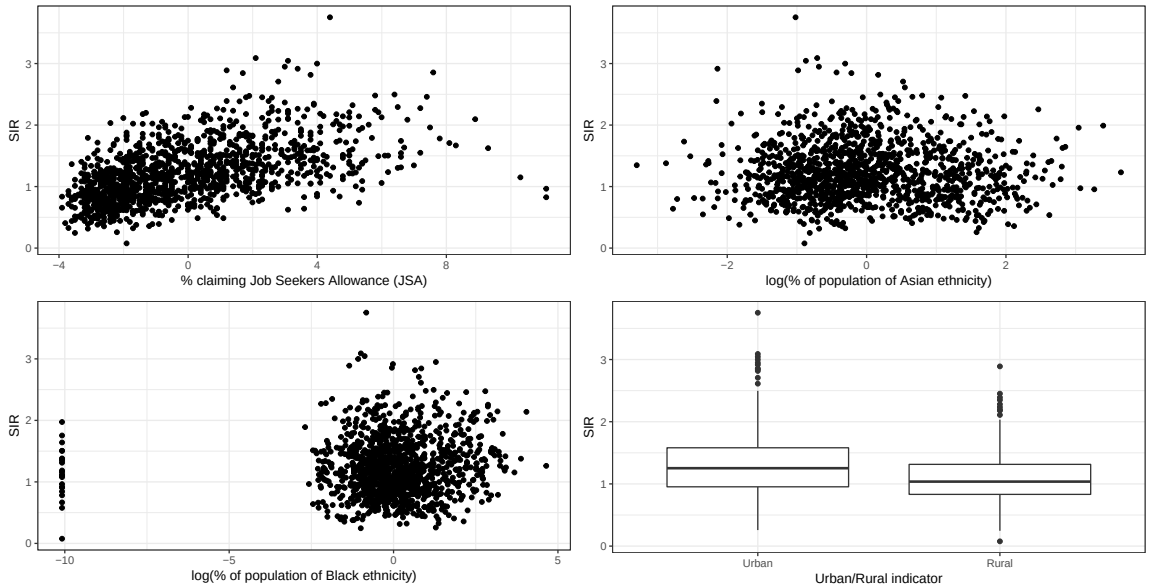

Figure 3.5: Scatterplots of the four potential covariates versus SIR. Top left: % claiming job seekers allowance. Top right: log(% of population of Asian ethnicity). Bottom left: log(% of population of Black ethnicity). Bottom right: Urban/rural indicator.

shows the relationship between SIR and the % of working age people claiming JSA in each of the IGs for all years. There appears to be a positive linear relationship, i.e. as the % claiming JSA increases, the SIR also increases. This indicates that areas where a high percentage of the population claim JSA could experience an increase in the risk of coronary heart disease. As previously stated, this relationship is to be expected as there is evidence that higher levels of deprivation in an area increases the risk of disease (Audit Scotland, 2012). The top right and bottom left plots show the relationship between SIR and the log % of population of Asian and Black ethnicity respectively, the natural logarithm of these covariates were taken as the data on the original scale was skewed. Neither of the plots show an obvious relationship with SIR, and so may not have any effect on the coronary heart disease risk of an area. Finally, the bottom right plot shows boxplots of SIR for areas classed to be urban and areas classed to be rural. The median SIR for urban areas is 0.992 whereas the median SIR for of rural areas is 0.833 suggesting that there might be a slight increase in risk of coronary heart disease for areas classed to be urban.

3.2.4 Exploratory analysis

In order to assess the presence of residual spatial correlation in the data, which needs to be accounted for, a Poisson generalised linear model (GLM) was fitted to

the data for the year 2003 with the covariates described before separately for each disease. Moran's I (Moran, 1950) statistics were then calculated using the residuals from the model, and the results show that strong spatial correlation was present after the covariate effects had been accounted for, with Moran's I statistics of 0.227 with significant associated p-value of < 0.001.

In order to assess the presence of temporal correlation, the average lag-one correlation coefficient was calculated for each disease across the IGs. However, given that we have a very short time series (only 10 time points) the results from this were inconsistent. Although, given that the population from which the data come from will remain broadly the same every year, *a priori*, we expect there to be temporal correlation and so we will account for this in the final model.

3.3 Methodology

Here we outline a Bayesian hierarchical model for these data with the aim of quantifying how health inequalities in coronary heart disease have changed over time in Scotland and at the health board level.

3.3.1 Likelihood model

The standard likelihood model typically used in this context is given by

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(e_{it}\theta_{it}), \quad i = 1, ..., n(= 1235); t = 1, ..., T(= 10), \\ \ln(\theta_{it}) &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \mathcal{H}_{h(i)t} + \phi_i, \quad h(i) = 1, \ldots, H(= 14), \end{aligned} \tag{3.1}$$

where Y_{it} and e_{it} are the observed and expected numbers of hospital admissions in IG i and time point t, while θ_{it} is the risk relative to the expected numbers e_{it}. We model the log-risk with 3 components, the first of which is the $p \times 1$ vector of known covariates $\boldsymbol{x}_i = (1, x_{i2}, ..., x_{ip})$, including an intercept term, with regression parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. Given that we do not have access to temporally-varying covariate information we cannot include this in the model. The prior specified is $\boldsymbol{\beta} \sim N(0, 100\boldsymbol{I})$ which is weakly informative to allow their values to be informed by

the data. The remaining two components are a baseline spatial effect ϕ_i, and a health board temporal trend $\mathcal{H}_{h(i)t}$, where $h(i)$ denotes that IG i is located within HB h. Both of these components are described in the following sections. We have chosen not to include temporal variation in the random effects ϕ_i, because we want all temporal variation to be incorporated in the temporally varying health board effects as these are of key interest.

3.3.2 Spatial effects

In Section 3.2.4, we found evidence of substantial residual spatial correlations in the data, which we model via spatial random effects. Spatial correlation is induced into these random effects via the neighbourhood matrix $\boldsymbol{W}$, which is an $(n \times n)$ binary matrix, where $w_{ij} = 1$ if two areas are defined to be neighbours and $w_{ij} = 0$ if not. Also $w_{ii} = 0$ for all i. There are many different ways to specify if two areas are neighbours, and the one used in this thesis is if two areas share a common border. Geographically, Scotland comprises the northern one third of Great Britain along with 790 surrounding islands of which only 130 are still inhabited by people. Therefore, there are several IGs which have no defined neighbours using this specification (6 in total). This presents a major problem when estimating the spatial random effects and so time was spent identifying these areas and connecting them to their closest neighbour using the euclidean distance. There are also several groups of IGs which are not connected to the mainland as the islands do not share a physical border with mainland Scotland and so it was decided to connect these IGs to the nearest mainland IG, again using euclidean distance. We model the spatial effects by a Leroux CAR prior (Leroux et al., 2000) given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}\right), \tag{3.2}$$

$$\tau^2 \sim \text{Inverse-Gamma}(0.001, 0.001),$$

$$\rho \sim \text{Unif}(0, 1),$$

where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. The parameter ρ controls the level of spatial autocorrelation in the data, with $\rho = 0$ corresponding to independence in space and $\rho = 1$ corresponding to the intrinsic CAR prior (Besag et al., 1991, see Section 2.5.1). The priors specified for hyperparameters τ^2 and ρ are weakly informative and allow their values to be informed mainly by the data. The conjugate inverse-gamma prior was used for τ^2 to allow this step to be implemented using Gibbs sampling.

3.3.3 Temporally varying HB effects

A key question in our analysis is to investigate the health inequalities between Scotland's 14 regional health boards, and how these change over time. Therefore, we include health board temporal trends in the model, $\boldsymbol{\mathcal{H}}_h = (\mathcal{H}_{h1}, ..., \mathcal{H}_{hT})$, which are modelled by the first-order autoregressive process

$$\begin{aligned} \mathcal{H}_{ht} &\sim N(\alpha \mathcal{H}_{h,t-1}, \sigma^2), \\ \sigma^2 &\sim \text{Inverse-Gamma}(0.001, 0.001), \\ \alpha &\sim \text{Unif}(0, 1), \end{aligned} \tag{3.3}$$

where $\mathcal{H}_{ht}$ is the effect for health board h and time point t. Temporal correlation is induced via the hyperparameter α, with $\alpha = 0$ indicating independence across time while $\alpha = 1$ indicates strong temporal dependence. A previous version of this model allowed the hyperparameters α and σ^2 to vary by health board, however in this case the parameters were not well identified by the data, and so a simpler prior with single parameters α and σ^2 was implemented instead. As before, weakly informative priors were assigned to α and σ^2 to allow their values to be mainly informed from the data, and both steps to be updated using Gibbs sampling.

3.4 Estimation

In order to obtain posterior summaries of each parameter, samples were drawn from the posterior distribution using Markov chain Monte-Carlo (McMC) simulation using

both Gibbs sampling and Metropolis steps. The McMC algorithm was written (as part of this thesis) in `R` (R Core Team, 2014). However, due to the large number of random effects that have to be sampled at each iteration, this step was implemented using the `R` package `Rcpp`, which allows for this script to be run in the more efficient language, `C++` (Eddelbuettel and François, 2011, Eddelbuettel, 2013). Given that the neighbourhood matrix $\boldsymbol{W}$ is a large but sparse matrix, we also utilised its triplet form to speed up computation. Details of each step of the McMC sampler for Model 3.1 are shown in the following section.

3.4.1 Update for $\boldsymbol{\beta}$

A Metropolis step is used to sample $\boldsymbol{\beta}$ given by

$$f(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X},\mathcal{H},\boldsymbol{\phi}) \propto \prod_{i=1}^{n}\prod_{t=1}^{T} \text{Poisson}(Y_{it}|\boldsymbol{x}_i,\boldsymbol{\beta},\mathcal{H},\boldsymbol{\phi}) \times \prod_{k=1}^{p} N(\beta_k|0,\sigma^2_\beta), \tag{3.4}$$

$$\begin{aligned}\ln(f(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X},\mathcal{H},\boldsymbol{\phi})) \propto &\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i) - \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i)\right)\\ &+\sum_{k=1}^{p}\left(-\frac{\beta_k^2}{2\sigma^2_\beta}\right),\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is drawn as a single block for all 4 covariates. Each of the continuous covariates were mean centered before being added to the model to remove correlation between the parameter estimates and the intercept and allow for easier interpretation of the HB effects.

3.4.2 Update for $\boldsymbol{\phi}$

Each ϕ_i is drawn separately using a Metropolis step as follows:

$$f(\phi_i|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\mathcal{H}) \propto \prod_{t=1}^{T} \text{Poisson}(Y_{it}|\boldsymbol{x}_i,\boldsymbol{\beta},\mathcal{H},\boldsymbol{\phi}) \times N(\phi_i|\boldsymbol{\phi_{-i}}), \tag{3.5}$$

$$\begin{aligned}\ln(f(\phi_i|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\mathcal{H})) \propto & \sum_{t=1}^{T}\left(y_{it}(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i)-\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i)\right) \\ & -\frac{1}{2\frac{\tau^2}{\rho\sum_{j=1}^{n}w_{ij}+(1-\rho)}}\left(\phi_i-\left(\frac{\rho\sum_{j=1}^{n}w_{ij}\phi_j}{\rho\sum_{j=1}^{n}w_{ij}+(1-\rho)}\right)\right)^2.\end{aligned}$$

Due to indentifiability issues, each ϕ_i was mean centered by health board.

3.4.3 Update for τ^2

τ^2 is drawn using a Gibbs sampler as follows:

$$\begin{aligned}f(\tau^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\mathcal{H},\boldsymbol{\phi}) &\propto N(\boldsymbol{0},\tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1}) \times \text{Inverse-Gamma}(a,b), \\ &\propto \text{Inverse-Gamma}(\tilde{a},\tilde{b}),\end{aligned} \tag{3.6}$$

where,

$\tilde{a} = a + \frac{n}{2}$,

$\tilde{b} = b + \frac{1}{2}\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\boldsymbol{W})\boldsymbol{\phi}$.

3.4.4 Update for ρ

Finally, ρ is drawn using a Metropolis step as follows:

$$\begin{aligned}f(\rho|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\mathcal{H},\boldsymbol{\phi}) &\propto N(\boldsymbol{0},\tau^2\boldsymbol{Q}(\rho,\boldsymbol{W})^{-1}), \\ \ln[f(\rho|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\mathcal{H},\boldsymbol{\phi})] &\propto \frac{1}{2}\sum_{i=1}^{n}\ln[\rho e_i+(1-\rho)] - \frac{1}{2}\frac{\boldsymbol{\phi}^\top\boldsymbol{Q}(\rho,\boldsymbol{W})\boldsymbol{\phi}}{\tau^2},\end{aligned} \tag{3.7}$$

where e_i is a vector of eigenvalues from the matrix $(\text{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W})$.

3.4.5 Update for $\mathcal{H}_h$

A Metropolis step is also used to sample the vector of $\boldsymbol{\mathcal{H}}$ effects, given by

$$f(\mathcal{H}_h|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\phi}) \propto \prod_{i=1}^{n}\prod_{t=1}^{T} \text{Poisson}(Y_{it}|\boldsymbol{x}_i,\boldsymbol{\beta},\mathcal{H}_h,\boldsymbol{\phi}) \times N(\mathcal{H}_h|\mathbf{0},\sigma^2\boldsymbol{R}^{-1}), \quad (3.8)$$

$$\begin{aligned}\ln(f(\mathcal{H}_h|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\phi})) \propto &\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i) - \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}+\mathcal{H}_{h(i)t}+\phi_i)\right)\\ &+\left(-\frac{\mathcal{H}_h^\top\boldsymbol{R}\mathcal{H}_h}{2\sigma^2}\right),\end{aligned}$$

where $\boldsymbol{\mathcal{H}}_h = (\mathcal{H}_{h1},...,\mathcal{H}_{hT})$ is updated in a block for all time points in a health board. There are therefore 14 separate updates (one for each health board) and within each block, 10 parameters are proposed. The HB effects were mean centered due to identifiability issues.

3.4.6 Update for σ^2

σ^2 is drawn using a Gibbs sampler as follows:

$$\begin{aligned}f(\sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\mathcal{H}},\boldsymbol{\phi}) &\propto \prod_{h=1}^{H} N(\mathbf{0},\sigma^2\boldsymbol{R}^{-1}) \times \text{Inverse-Gamma}(c,d), \qquad (3.9)\\ &\propto \text{Inverse-Gamma}(\tilde{c},\tilde{d}),\end{aligned}$$

where,

$\tilde{c} = c + \frac{HT}{2}$,

$\tilde{d} = d + \frac{1}{2}\sum_{h=1}^{H}\mathcal{H}_h^\top\boldsymbol{R}\mathcal{H}_h$.

3.4.7 Update for α

α is drawn using a Gibbs sampler as follows:

$$\begin{aligned}f(\alpha|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\mathcal{H}},\boldsymbol{\phi}) &\propto \prod_{h=1}^{H} N(\mathbf{0},\sigma^2\boldsymbol{R}^{-1}) \times \text{Unif}(0,1), \qquad (3.10)\\ &\propto N(\mu,\tilde{\sigma}^2),\end{aligned}$$

	Posterior Median	95% CI
Spatial Autocorrelation	0.437	(0.325, 0.560)
Temporal Autocorrelation	0.875	(0.810,0.962)

Table 3.1: Estimates and 95% credible intervals for autocorrelation in model.

where,

$$\mu = \frac{\sum_{h=1}^{H}\sum_{t=2}^{T}\mathcal{H}_{h,t}\mathcal{H}_{h,t-1}}{\sum_{h=1}^{H}\sum_{t=2}^{T}\mathcal{H}_{h,t-1}^2},$$
$$\tilde{\sigma}^2 = \frac{\sigma^2}{\sum_{h=1}^{H}\sum_{t=2}^{T}\mathcal{H}_{h,t-1}^2}.$$

3.5 Results

The spatio-temporal model proposed in Section 3.3 was applied to the data for Scotland described in Section 3.2. Inference is based on 150,000 McMC samples with a burn-in period of 50,000. The chain was thinned by 5, due to limitations in computer memory and to make the samples closer to independent, and so the posterior estimates are based on a total of 20,000 samples. Convergence was checked both by examining parameter trace plots and Geweke diagnostics (Geweke, 1992).

3.5.1 Spatial and Temporal Autocorrelation

Table 4.1 shows the posterior medians and 95% credible intervals for the autocorrelation parameters in the model. First of all, the estimate for the spatial autocorrelation parameter ρ is 0.437, this indicates there is moderate spatial autocorrelation in the data. The temporal autocorrelation, α is estimated to be 0.875, which suggests reasonably high temporal autocorrelation within the health boards.

3.5.2 Health board effects

In order to investigate whether there are health inequalities between Scotland's 14 health boards and how these are changing over time (Question 1, Section 3.1), Figure 3.6 shows the posterior medians for each HB on the risk scale, $\theta_{ht} = \exp(\mathcal{H}_{ht})$, for coronary heart disease, after adjusting for the known covariates. It can be seen from this plot that there are health inequalities between the HBs as there are differences between the estimated HB posterior medians for coronary heart disease. The risk of

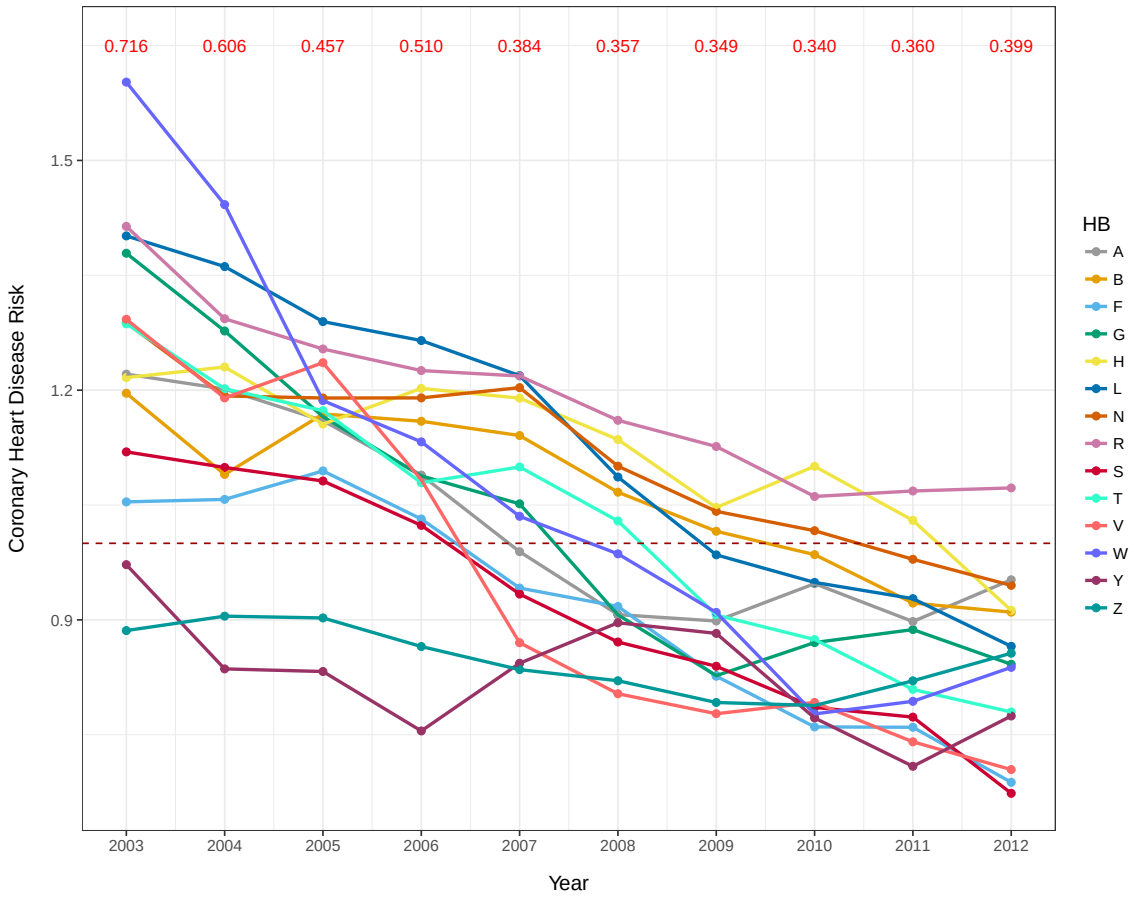

Figure 3.6: Health board risk effects across time ($\theta_{ht} = \exp(\mathcal{H}_{ht})$). Posterior medians shown for all health boards. The numbers at the top of each graph represent the range in the median HB effects for each year.

disease is not consistent between health boards, nor is it constant over time. However, these inequalities are decreasing over time, with a difference of 0.716 between the highest and lowest median HB risk in 2003 compared to 0.399 in 2012. Given that the island boards are significantly smaller than the mainland boards we tend to see greater variation in risk estimates for these boards over the time period. In fact, in 2003 the HB with the highest posterior median is Western Isles (W) and the HB with the lowest posterior median is Shetland (Z), both of which are islands. It could, therefore, be the case that the narrowing in health inequality between the health boards is being driven by the highly variable island HBs. However, if these HBs are removed the range in posterior medians in 2003 is 0.421 compared to 0.279 in 2012 so, even when ignoring the island boards, a reduction is still seen in health inequality over time. Therefore, although there are still differences in the HB effects at the end of the time period, it appears that these inequalities have reduced slightly from 2003 to 2012.

Figures 3.7 and 3.8 show the overall health board effects for each health board

over time. These plots show the risk effects for the IGs belonging to each of the 14 health boards after the covariate effects have been removed. One of the main features of these plots is that compared to the boxplots of the SIRs for each HB over time, shown in Figure 3.2, the median lines are much smoother. This to be expected given that the goal of a statistical model is to estimate the underlying trend in the data rather than capture random noise and induced temporal autocorrelation. In general, most HBs show a decreasing trend, as expected from the raw data, however there are still differences in the trends between the HBs. For example, some of the HBs have a decreasing trend at the beginning of the time period which then levels out and remains reasonably constant at the end of the time period, e.g. Ayrshire and Arran (A), Greater Glasgow and Clyde (G), Forth Valley (V) and Western Isles (W). Whereas others show little change at the start of the time period and then begin to decrease towards the end, e.g. Borders (B), Fife (F), Highland (H) and Grampian (N). Some decrease fairly consistently over the entire period, e.g. Lanarkshire (L), Orkney (R), Lothian (S) and Tayside (T). Finally, there are two boards whose trends do not follow the general decreasing pattern, namely, Dumfries and Galloway (Y) and Shetland (Z). For both these boards, the coronary heart disease risk remains slightly below the null risk line of one for most time periods, which shows that, unlike the other boards, the coronary heart disease risk effect associated with these boards is low throughout the entire study period. There are also differences in the magnitude of the change in trend. For example, there are some HBs whose risk at the start of the time period is much higher, e.g. Greater Glasgow and Clyde (G), Lanarkshire (L), Orkney (R), Tayside (T), Forth Valley (V) and Western Isles (W), all have high median coronary heart disease risk estimates in 2003. By the end of the time period, all of these HBs except from Orkney, have median risk plus 95% credible intervals below the null risk line of 1 and so the change in coronary heart disease risk over time is the most extreme for these boards. Finally, it should be noted that the uncertainty around these estimates for some of the smaller boards (particularly the island boards) is greater than for the larger HBs. This is to be expected given that the boards with small numbers of IGs have much less data to estimate these effects. In general, although these plots show decreasing trends across each of the 14 health

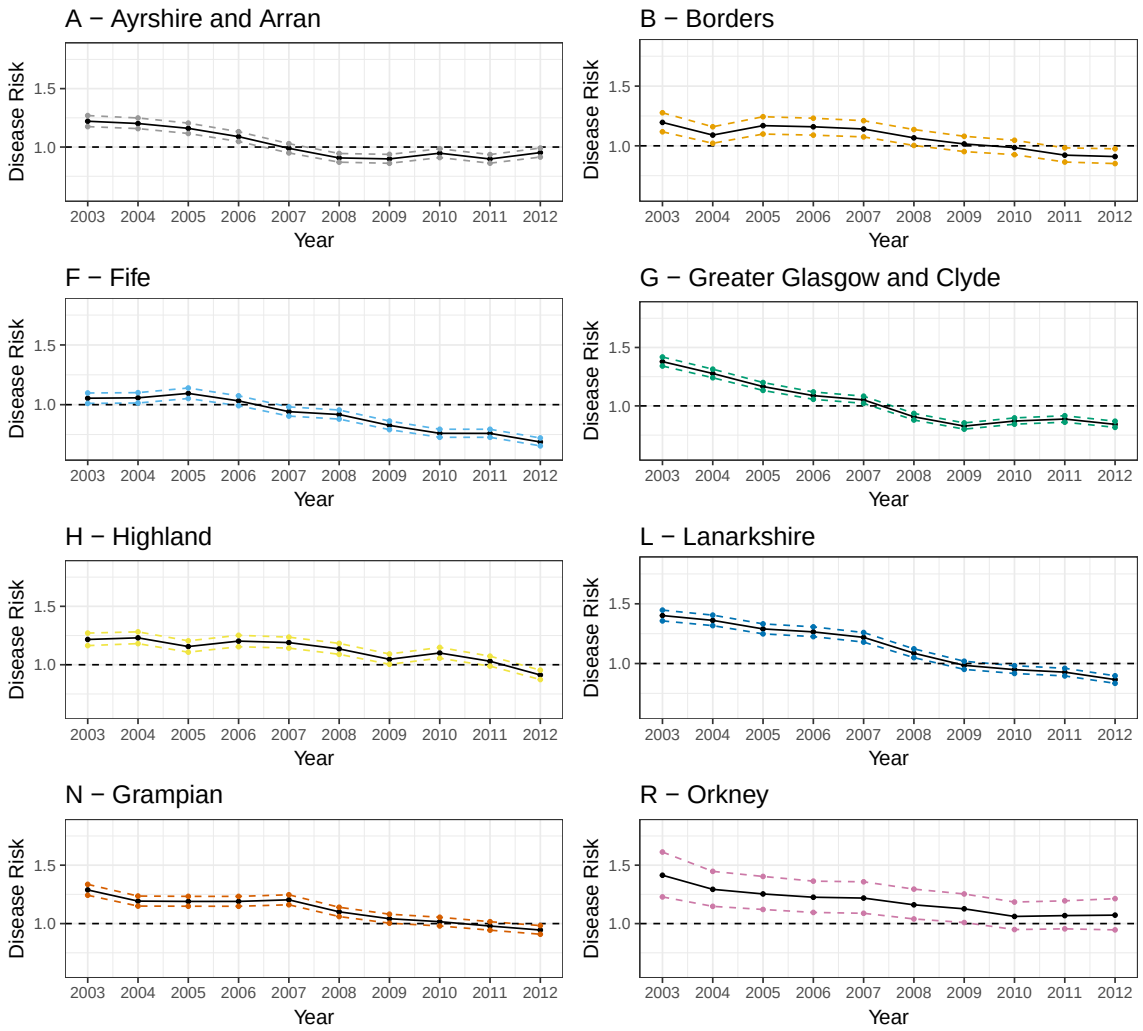

Figure 3.7: Health board risk effects across time ($\theta_{ht} = \exp(\mathcal{H}_{ht})$). Posterior medians in black with 95% credible intervals shown by coloured dashed bands. Black dashed line indicates risk of 1.

boards, the shape of these trends is not consistent across HB. This indicates that even after the covariate effects have been removed, there still appear to be differences in coronary heart disease risk due to which health board each IG is located in.

3.5.3 Risk Maps

The risk estimates (posterior medians) are shown in Figure 3.9 for the years 2003, 2006, 2009 and 2012. A clear pattern can be seen with coronary heart disease risk decreasing over the time period. In 2003, there are many areas in Scotland with increased risk estimates, which can be seen clearly as the areas shaded in red. In 2006, although there are still some areas with increased risk of coronary heart disease, the number has decreased. For the maps of 2009 and 2012, we now see that most areas are shaded in blue which suggests that for most areas in Scotland, the risk of

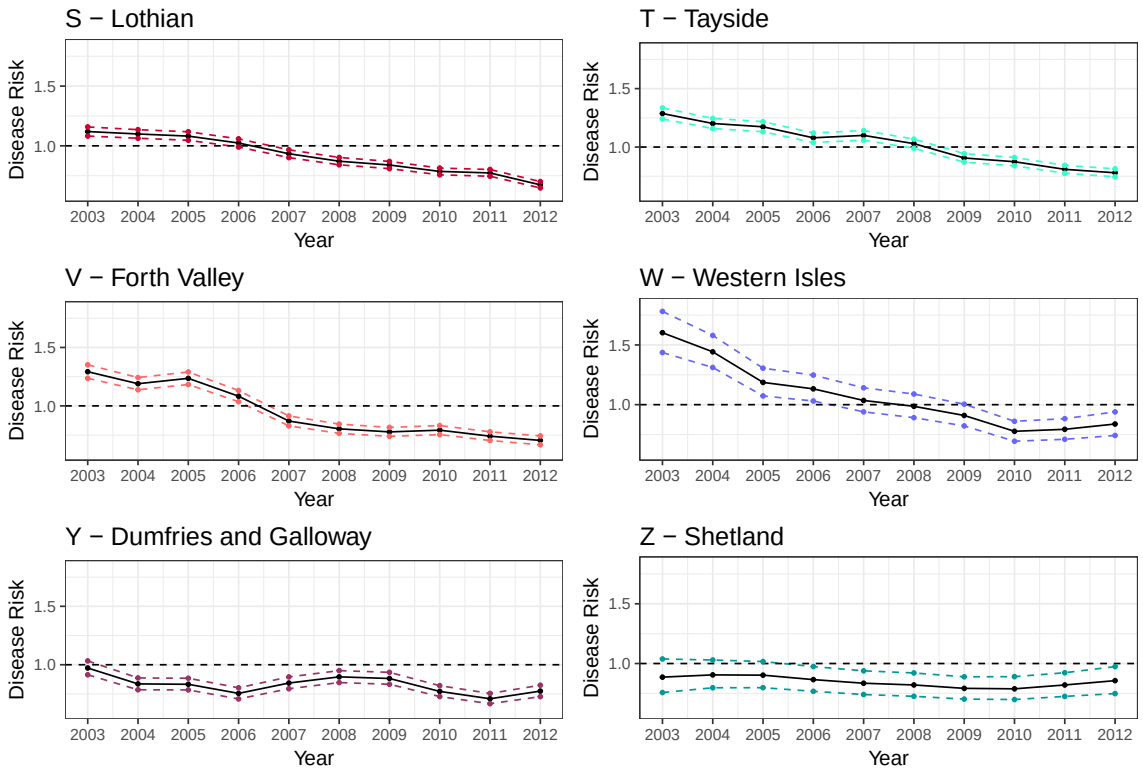

Figure 3.8: Health board risk effects across time ($\theta_{ht} = \exp(\mathcal{H}_{ht})$). Posterior medians in blackwith 95% credible intervals shown by coloured dashed bands. Black dashed line indicates risk of 1.

coronary heart disease has decreased over time. When these risk maps are compared to the maps of SIR in Figure 3.3 it should be noted that the estimated risk maps are smoother than the raw SIR values. This is to be expected given the nature of spatial smoothing where random effects borrow strength from their neighbours. Modelled in this way the chance of extreme risks occuring is reduced. For example, over all IGs and time points, the SIR for coronary heart disease ranges from 0.00 to 3.75, while the corresponding model risk estimates range between 0.20 and 2.20.

The significance of disease risk for the same years is shown in Figure 3.10. Areas shaded in blue have significantly lower disease risks than average (credible intervals for θ_{it} that are less than 1), areas shaded in red have disease risks that are significantly higher than average (credible intervals above 1) and areas in grey have credible intervals that contain 1 and therefore show no significant difference in risk on average. In 2003, only 14% of areas have significantly decreased risk of coronary heart disease, 33% contain the null risk of 1 and the majority of areas, 53%, have increased risk effects. Compare this to 2012 where 74% of areas have significantly decreased risk effects and only 4% of areas have significantly increased risk effects. From this we

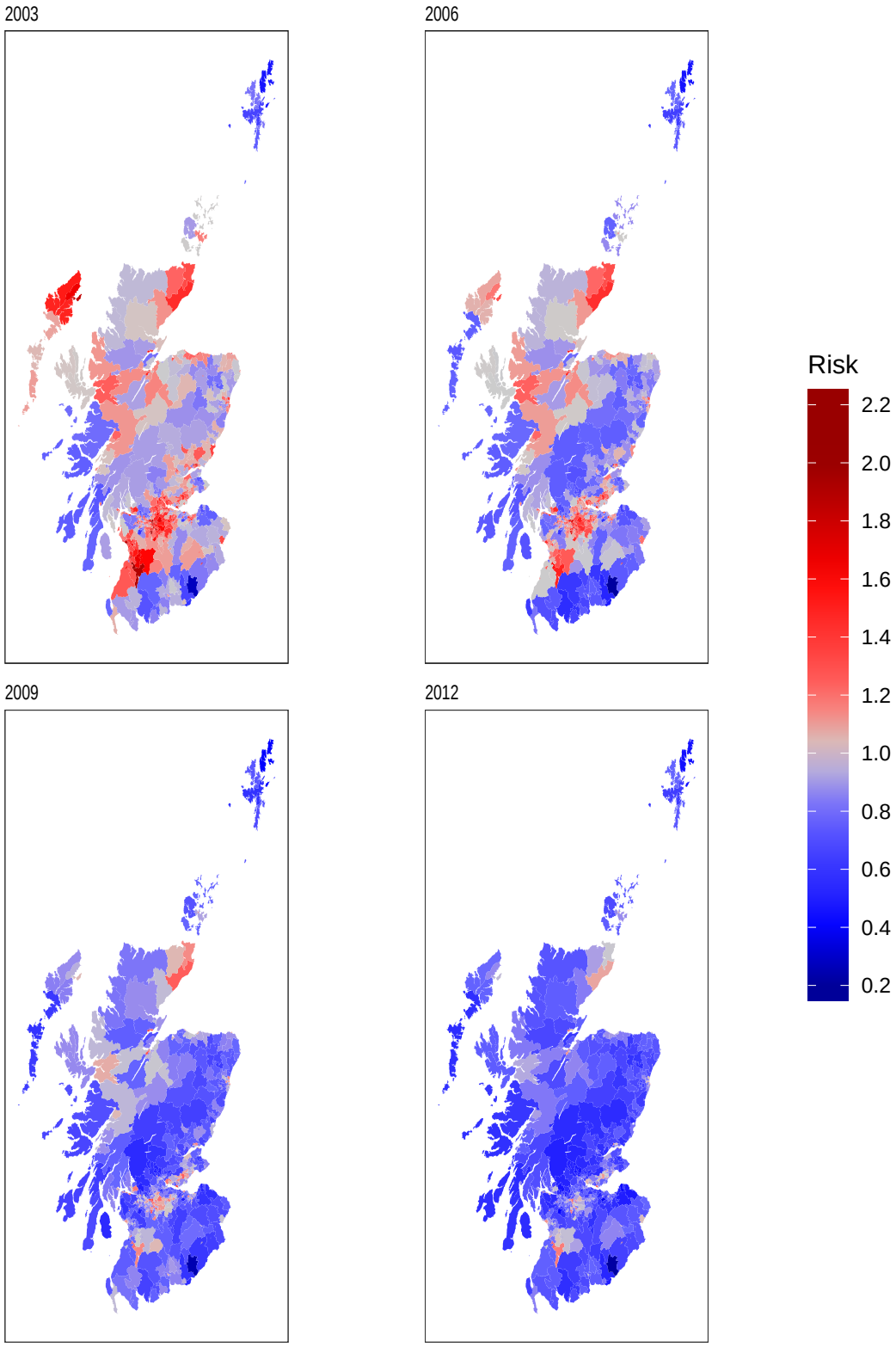

Figure 3.9: Risk estimates for coronary heart disease in IGs in Scotland in 2003, 2006, 2009 and 2012.

can see huge changes in the risk of coronary heart disease in Scotland, with around 60% more areas exhibiting substantially decreased risk of coronary heart disease in 2012 compared to 2003.

3.5.4 Overall health inequalities

In order to investigate whether the overall health inequalities have changed over time across the IGs in Scotland (and not just between HBs as in Section 3.5.2) (Question 2, Section 3.1), Figure 3.11 shows boxplots of the posterior median disease risk for all IGs from 2003 to 2012. Most obviously from this plot, we see an overall decreasing trend in the risk of coronary heart disease in Scotland from 2003 to 2012.

However, in order to assess changes in health inequalities, the variation in coronary heart disease risk should be looked at rather than the median level. This will give an indication into the difference between the areas with the lowest risk of coronary heart disease and the areas with the highest risk. A reduction in health inequality can be assessed either by looking for a narrowing in the width of the boxplots or by a decrease in the interquartile range (IQR) which is printed in red above each boxplot. It is quite clear from these plots that, not only is overall risk in coronary heart disease going down in Scotland over time, but the inequality in coronary heart disease risk is also decreasing, as the IQRs decrease year on year. This indicates that the differences in coronary heart disease risk between population areas has reduced from 2003 to 2012. It should also be noted that, not only are the boxplots narrower, but the tails also reduce in length over time and the risk levels associated with the outliers are much lower in 2012 compared with 2003. This tells us that we can also see a decrease in risk in the areas which are the most at risk of coronary heart disease and that the differences between the areas in the extremes of the risk levels are much smaller in 2012. In fact, in 2003, the area with the highest risk of coronary heart disease had a estimate of 2.187, i.e. this area was more than two times more at risk of coronary heart disease than average, whereas in 2012 this estimate was reduced to 1.567. This is important as it shows that not only is coronary heart disease risk decreasing for the areas who had low or average levels of coronary heart disease risk to begin with, but we are seeing this decrease in coronary heart disease risk over all

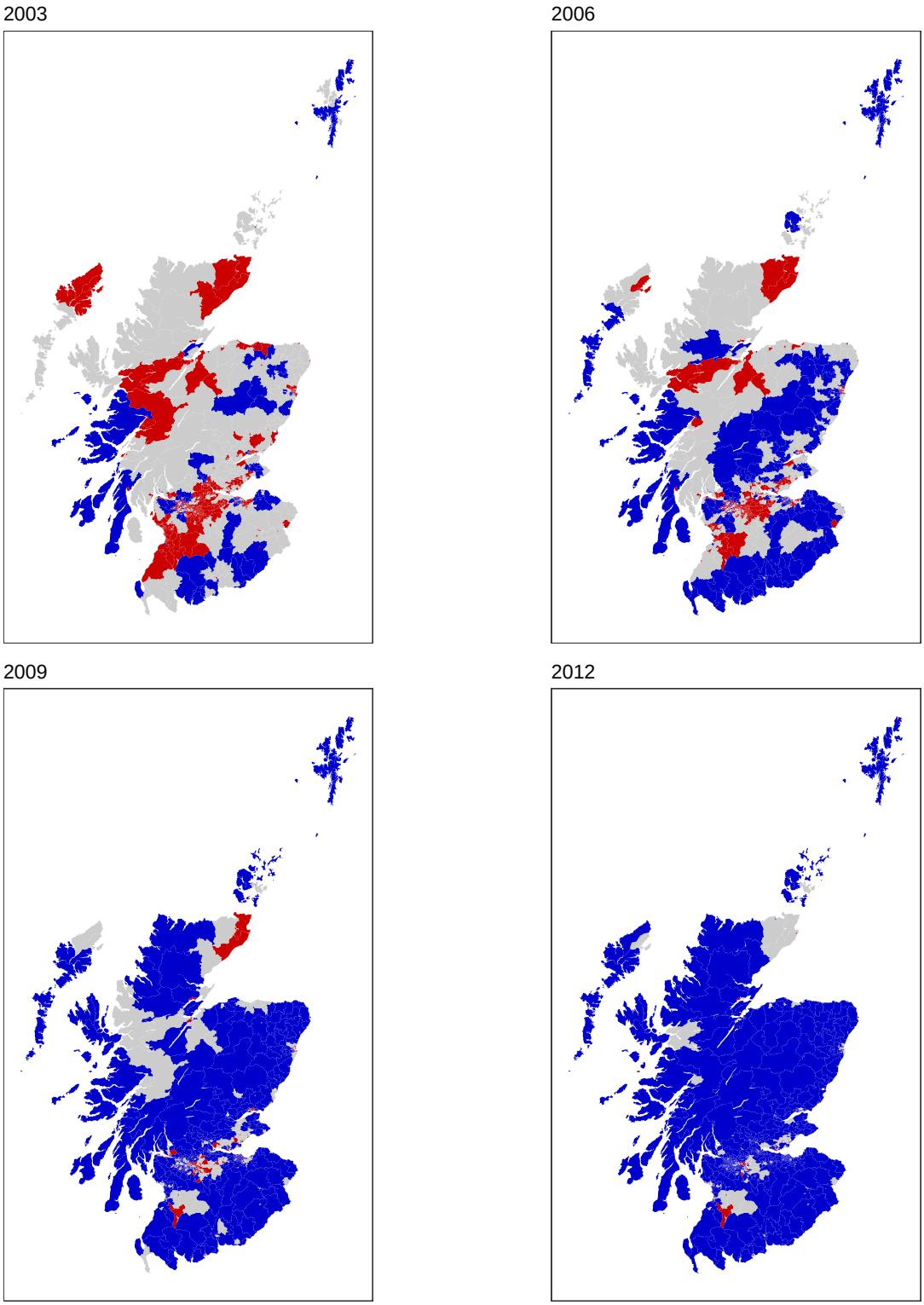

Figure 3.10: Significance of the risk estimates for coronary heart disease in IGs in Scotland in 2003, 2006, 2009 and 2012. Areas shaded in blue have significantly lower disease risks (credible intervals for θ_{it} that are less than 1), areas in grey have credible intervals that contain 1 and areas shaded in red have disease risks that are significantly higher than average.

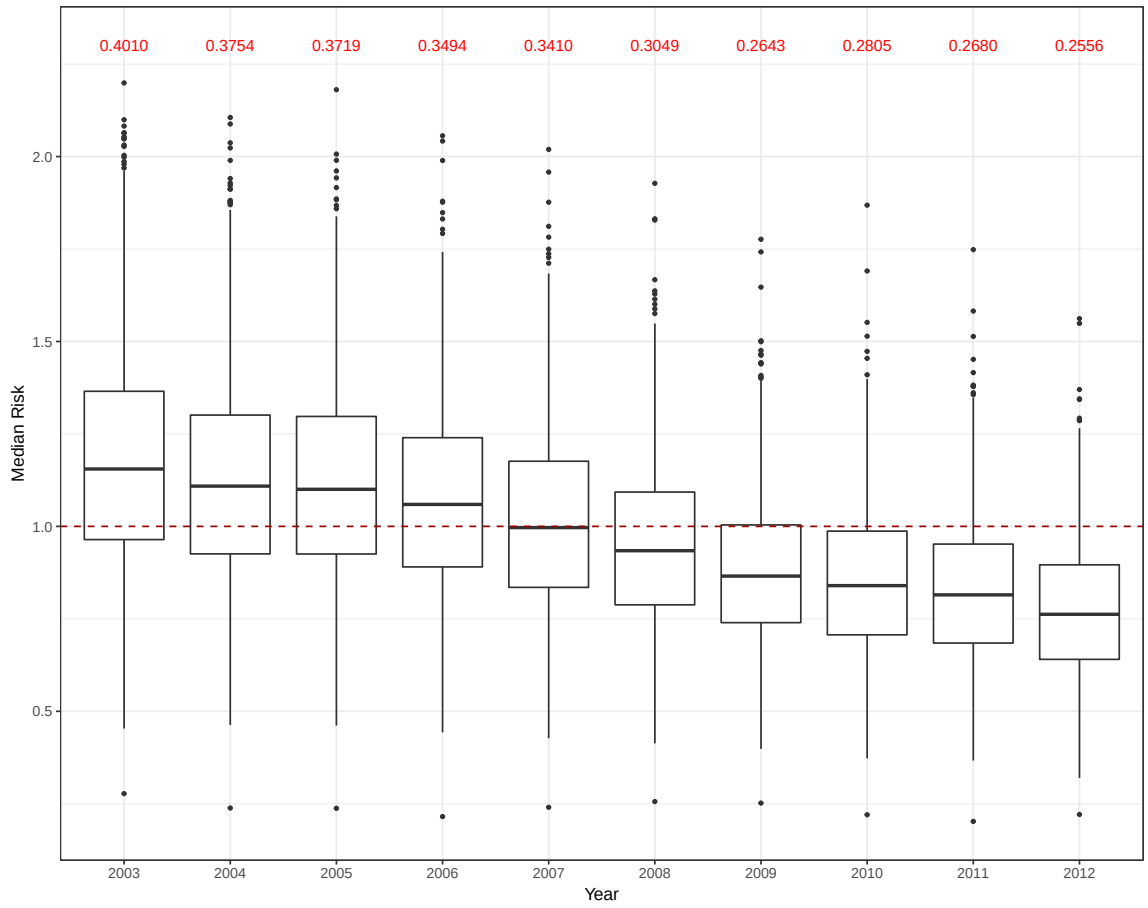

Figure 3.11: Boxplots of risk for coronary heart disease in IGs in Scotland from 2003 - 2012. The IQR across IGs are printed in red. Outliers are those observations that lie outside 1.5(IQR)

Covariate	Median RR	95% CI
% 16-64 claiming JSA	1.062	(1.058, 1.067)
Log % Asian	0.985	(0.971, 0.999)
Log % Black	1.008	(1.002, 1.014)
Rural area	0.956	(0.929, 0.984)

Table 3.2: Relative risk estimates for a 1% increase in each covariate (not urban/rural covariate) and 95% credible intervals for the covariates in model.

the IGs and, crucially, Scotland's most vulnerable population.

3.5.5 Covariate effects

The effects of the covariates are displayed in Table 3.2. Presented are the estimates (posterior medians) and 95% credible intervals on the relative risk scale (Question 3, Section 3.1). The median relative risk (RR) for % 16-64 people claiming job seekers allowance is around 1.062 for a 1% increase, so the risk of coronary heart disease in an IG increases by 6.1% as the % claiming JSA increases by 1%. Since this is the proxy measure of deprivation included in the model, it can be inferred that as deprivation level increases, the risk of coronary heart disease also increases. The

median RR estimate for log(% of population of Asian ethnicity) is 0.985, suggesting that there may be a very small decrease in coronary heart disease risk as this covariate increases. This result seems in line with findings from a Scottish Government report (The Scottish Government, 2012) which found that those of Chinese ethnicity were the least likely to be diagnosed with cardiovascular disease (which includes all diseases of the heart and circulation including coronary heart disease) compared to the national average. In that report, those in Indian and Pakistani ethnic groups showed no difference compared to the national average. In this thesis, % of population of Asian ethnicity includes all Asian ethnic groups, which could explain the small protective effect for this covariate. The median RR estimate for log(% of population of Black ethnicity) is 1.008 and the 95% interval is entirely above 1, although the lower bound is very close to 1. This suggests a small increase in risk of 1% coronary heart disease as log(% of population of Black ethnicity) increases by 1%. Although most reports tend to find that those of Black ethnicity have lower risk of coronary heart disease, the result here could be due to the fact that the % of population of Black ethnicity is very low, the median value across all areas is only 0.25% and so there is very little data to estimate this parameter reliably. Finally, the 95% credible interval for rural areas compared to urban areas is entirely below 1 suggesting that areas which are rural are likely to have a decreased risk of coronary heart disease compared to urban areas. The risk associated with urban areas compared with rural areas is $\frac{1}{0.956} = 1.046$, i.e. there is an estimated increased risk of coronary heart disease of 4.6% when living in an urban area compared with a rural area. This result is in line with findings from recent Scottish Government publication (The Scottish Government, 2015) which states that the overall health of those living in rural areas is better compared to urban areas, with male life expectancy being nearly 3 years higher for rural areas than the rest of Scotland and the female life expectancy being nearly 2 years higher.

3.6 Discussion

In this chapter a hierarchical Bayesian model has been proposed to model the risk of disease across Scotland and over time. The model included covariate effects, spatial

random effects to allow for residual variation in space across the study area, and health board effects which are allowed to vary over time. This model was applied to hospital admissions data for coronary heart disease collected from 2003 to 2012 at intermediate geography level.

Overall we have found that there are differences in coronary heart disease risk across Scotland and these risks are also changing over time, with the median overall risk decreasing from 1.156 in 2003 to 0.762 in 2012. These differences in risk between population areas across Scotland are partly due to the covariates with increased levels of the percentage of the population claiming job seekers allowance (a measure of deprivation) inflating the risk of coronary heart disease by 6.1%, along with the percentage of the population of Black ethnicity by 1.1%, however this covariate may not be particularly well estimated given the very low percentages in the data, and the percentage of the population of Asian ethnicity decreasing risk. It was also found that living in a rural area compared to an urban area may have very small protective effect on coronary heart disease risk.

Although looking for overall trends in disease risk is important, it is also crucial to study the inequalities in disease risk, as often although an overall decrease is seen in disease risk, the areas which are most at risk do not necessarily follow this trend and so the inequality in disease actually increases. Here, however, we found that the inequality in risk is also decreasing, with the IQR of risk estimates for IGs in each year decreasing from 0.401 in 2003 to 0.256 in 2012, as shown in Figure 3.11. Given the reasonably short time period used here, these reductions seem fairly considerable given coronary heart disease is a chronic disease.

It has also been found that after adjusting for deprivation (and other covariates), health inequalities in coronary heart disease risk still exist between the health boards, although these have decreased over time. This can be seen most clearly from Figure 3.6, with the difference between the HB with the highest posterior median and lowest posterior median being 0.716 in 2003 and 0.399 in 2012. Although all 14 HBs show a decrease in coronary heart disease risk, the extent of this decrease varies hugely. The HB which shows the largest decrease is Western Isles (W) whose posterior median reduced by 0.764 over the 10-year period. Greater Glasgow and Clyde (G), Lanark-

shire (L), Tayside (T) and Forth Valley (V) all showed a decrease of greater than 0.5 from 2003 and 2012. For the two HBs whose posterior median was below the null risk of 1 in 2003, Dumfries and Galloway (Y) and Shetland (Z), although these HBs showed a small decrease in risk effect, the extent was nowhere near as large as the rest, which is to be expected, as they had less room to improve.

We were also interested in whether or not the smoking ban which came into effect in 2006 or the set up of a Ministerial Task Force to tackle health inequalities in 2007 had any effect on the results. However, there is no compelling evidence to show that either of these have made a significant impact since the reduction in risk of coronary heart disease started before the ban and Taskforce were implemented. However, data is only available until 2012 and these initiatives may not have had time to impact coronary heart disease risk yet. A further study over a longer time period may be required to investigate any potential lagged effects of the introduction of both the smoking ban and the Ministerial Task Force.

This methodology allows us to compare health inequalities in Scotland overall and between the 14 regional health boards for a single disease, in this case coronary heart disease. Although this same approach could be adopted for other diseases, investigating just one disease at a time ignores any correlation between diseases and will give an incomplete picture of overall inequality. Therefore, in Chapter 4 we propose a novel multivariate spatio-temporal model to quantify health inequalities in Scotland across 3 major diseases, which will enable us to better understand how they have changed over time. A multivariate approach is also beneficial as it will allow the model to borrow strength not only between neighbouring areas and time points as in the model proposed in this chapter, but also between disease.

Chapter 4

A multivariate model for estimating the changes in health inequalities across Scotland over time

4.1 Introduction

Measuring a population's health and the inequalities between population areas is an extremely complex problem. Therefore, the simplistic approach in Chapter 3, where a single disease is used to estimate health inequality does not seem adequate. This chapter extends this simple, univariate disease risk model to a more realistic multivariate disease risk model, where multiple diseases are investigated to better understand how health inequalities have changed across Scotland during the time period 2003 - 2012, using data containing hospital admissions for two more of Scotland's biggest killers (Scotpho, 2016), namely cerebrovascular disease and respiratory disease. Studying these diseases simultaneously with coronary heart disease will provide a better understanding of the relationships that exist between them.

As discussed in Section 2, very few multivarate space-time models have been proposed for modelling the risk of multiple diseases in space and time simultaneously

and so this chapter proposes a novel spatio-temporal multi-disease model for quantifying health inequalities in Scotland. The main focus is answering several questions of interest:

1. Are there health inequalities between Scotland's health boards and how are these changing over time?
2. Within a health board, how do average risk levels and temporal trends change between diseases?
3. How are health inequalities changing over time in IGs in Scotland across multiple diseases?
4. What impact do the covariates have on risk and how does this change by disease?
5. Are there some areas which have high risk for all three diseases?

We will present the results from our study and answer these questions of interest in Section 4.5. However, first the data are presented in Section 4.2, while our proposed model, which is a multivariate extension of the model proposed in Chapter 3, is presented in Section 4.3. Finally, Section 4.6 provides a discussion on the conclusions drawn from this study and possible ways in which it could be developed.

4.2 Data

Similar data as described in Chapter 3 is also held for cerebrovascular disease and respiratory disease. These data are the yearly counts of the numbers of hospital admissions for the period 2003 to 2012. For each year and IG we have the number of admissions to non-psychiatric/non-obstetric hospitals in Scotland with a main diagnosis of each disease for both sexes and all ages combined. Cerebrovascular disease is defined using the International Classification of Diseases Volume 10 (ICD10) codes (I60:I69, G45), while the codes are (J00:J99, R09.1) for respiratory disease. Expected numbers for both diseases were also calculated using indirect standardisation,

using age and sex adjusted rates for the year 2006/07, which were obtained from the Information and Services Division of the NHS.

For consistency, the same covariates used in Chapter 3 are included here, which are, the percentage of 16-64 year old's claiming job seekers allowance (JSA), the percentage of the population of Asian ethnicity, the percentage of the population of Black ethnicity and an urban/rural factor.

4.2.1 Exploratory analysis

The simplest measure of disease risk, the SIR, is a good way to informally explore risk patterns in the data. This section will provide some exploratory insight into the inequality in risk for coronary heart disease, respiratory disease and cerebrovascular disease.

Figure 4.1 shows boxplots of SIR by year and by health board for cerebrovascular disease, coronary heart disease and respiratory disease. For all three diseases, different patterns over time can be seen. As seen in Chapter 3, a decreasing trend can be seen over time for coronary heart disease. A decreasing trend can also be seen for cerebrovascular disease, however, it is much more gradual than that for coronary heart disease. Conversely, for respiratory disease, an increasing trend can be seen over the time period, suggesting that risk of respiratory disease is getting worse in Scotland over these years. When SIR is split by health board, some variation between the HBs can be seen. For each disease, there are some HBs whose median line sits above the null risk line and some whose median line sits below this line. For example for respiratory disease, Ayrshire and Arran (A), Greater Glasgow and Clyde (G) and Lanarkshire (L) all have median lines above the null risk line, and the rest of the HBs have median lines below the null risk line. This indicates that there may be health inequalities between the HBs within each disease. When comparing these plots between disease, there are some similarities, e.g. the median risk for Borders (B), Fife (F), Highland (H), Grampian (N), Orkney (R), Lothian (S), Dumfries and Galloway (Y) and Shetland (Z) is clearly below the null risk line for all three diseases, suggesting these boards have lower risk for all diseases on average over the 10 years. However, there are also some HBs where the plots show some discrepancies across

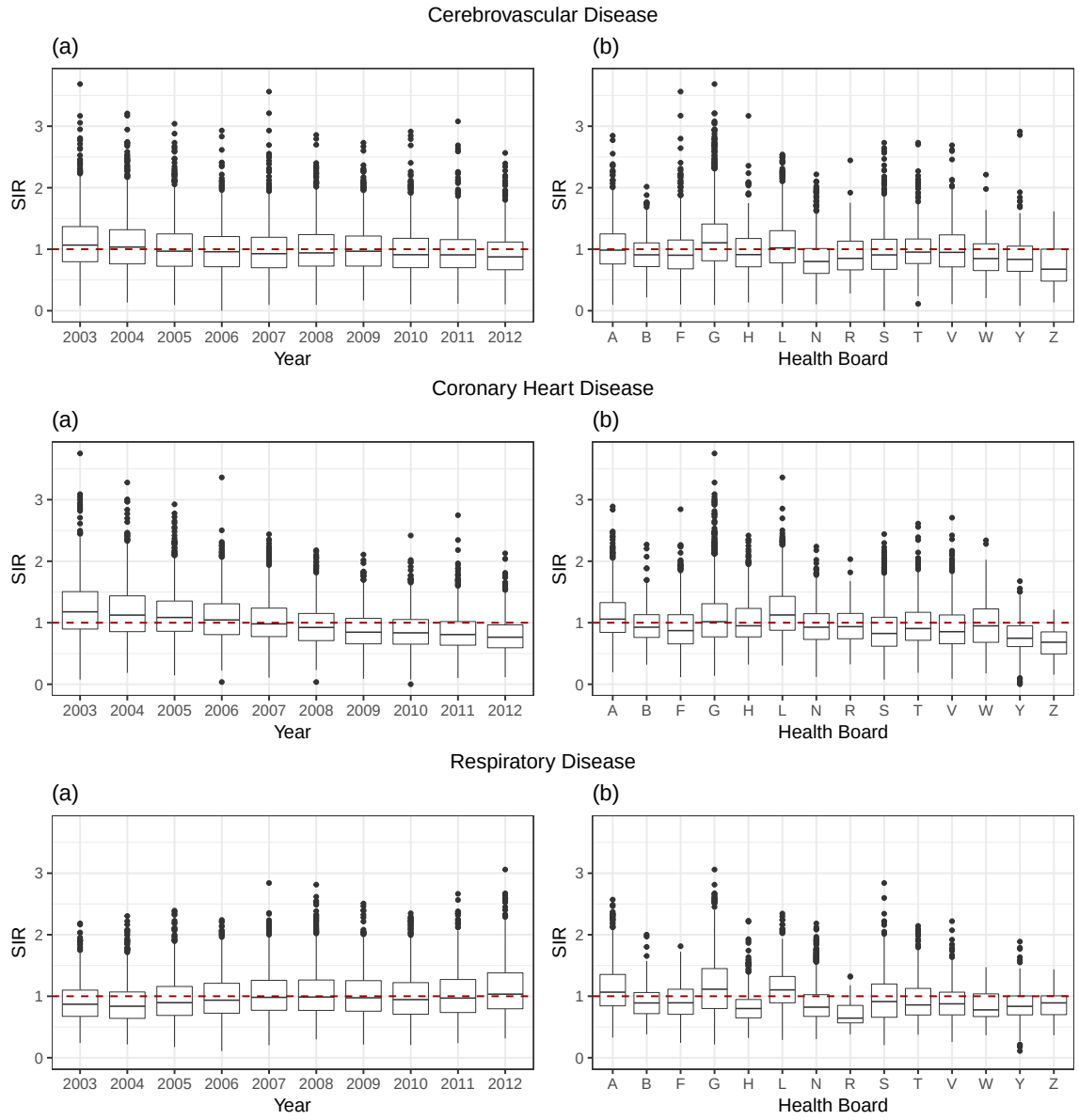

Figure 4.1: (a) Boxplots of the standardised incidence ratio (SIR) for cerebrovascular, coronary heart disease and respiratory disease admissions for IGs in Scotland from 2003 to 2012 by year. (b) Boxplots of the SIR for cerebrovascular, coronary heart disease and respiratory disease admissions for IGs in Scotland from 2003 to 2012 by health board. Red dashed line indicates a risk of 1.

the three diseases. For example, for Greater Glasgow and Clyde (G), the median is clearly above 1 for respiratory disease and cerebrovascular disease, but on the line for coronary heart disease. Similarly for Lanarkshire (L), the median line is above 1 for two of the three diseases (coronary heart disease and respiratory disease) but on the line for one (cerebrovascular disease). This shows that although for some HBs the patterns in risk may be similar across disease, this is not necessarily the case for all HBs.

To allow us to investigate how disease risk changes over time within HBs and whether these changes are consistent across HB and over the three diseases, Figure 4.2 shows boxplots of SIR split by HB and year for each disease. The first thing to note is that similar trends can be seen across some of the HBs in each disease, for example many of the HBs for cerebrovascular disease and coronary heart disease have decreasing trends, whereas for respiratory disease many of the HBs have an increasing trend. However, not all of the HBs within a disease follow these general patterns, e.g for respiratory disease Highland (H), Grampian (N), Forth Valley (V), and Dumfries and Galloway (Y) all seem to have reasonably constant risks over the time period, whereas the rest of the HBs seem to have increasing SIR values over time. As touched on previously, an important feature of these plots is that, when comparing each HB across the three diseases, the patterns shown are not always similar. For example when looking at Greater Glasgow and Clyde (G) the trend over time is strong and decreasing for coronary heart disease, strong and increasing for respiratory disease and although decreasing again for cerebrovascular disease, the extent of this decrease is much more gradual than for coronary heart diease. This is a good example of why looking at more than one disease is important in this situation, since there are some differences in the relationships observed across disease which may prove to be important when drawing conclusions about how health inequalities have changed across Scotland over time. Finally, the plots show much more variability over time for the three island HBs, Orkney (R), Western Isles (W) and Shetland (Z), due to the small number of IGs in these HBs.

In order the assess the presence of spatial variation in the data for each disease, Figure 4.3 shows the SIRs across IGs in Scotland in 2006 for cerebrovascular disease,

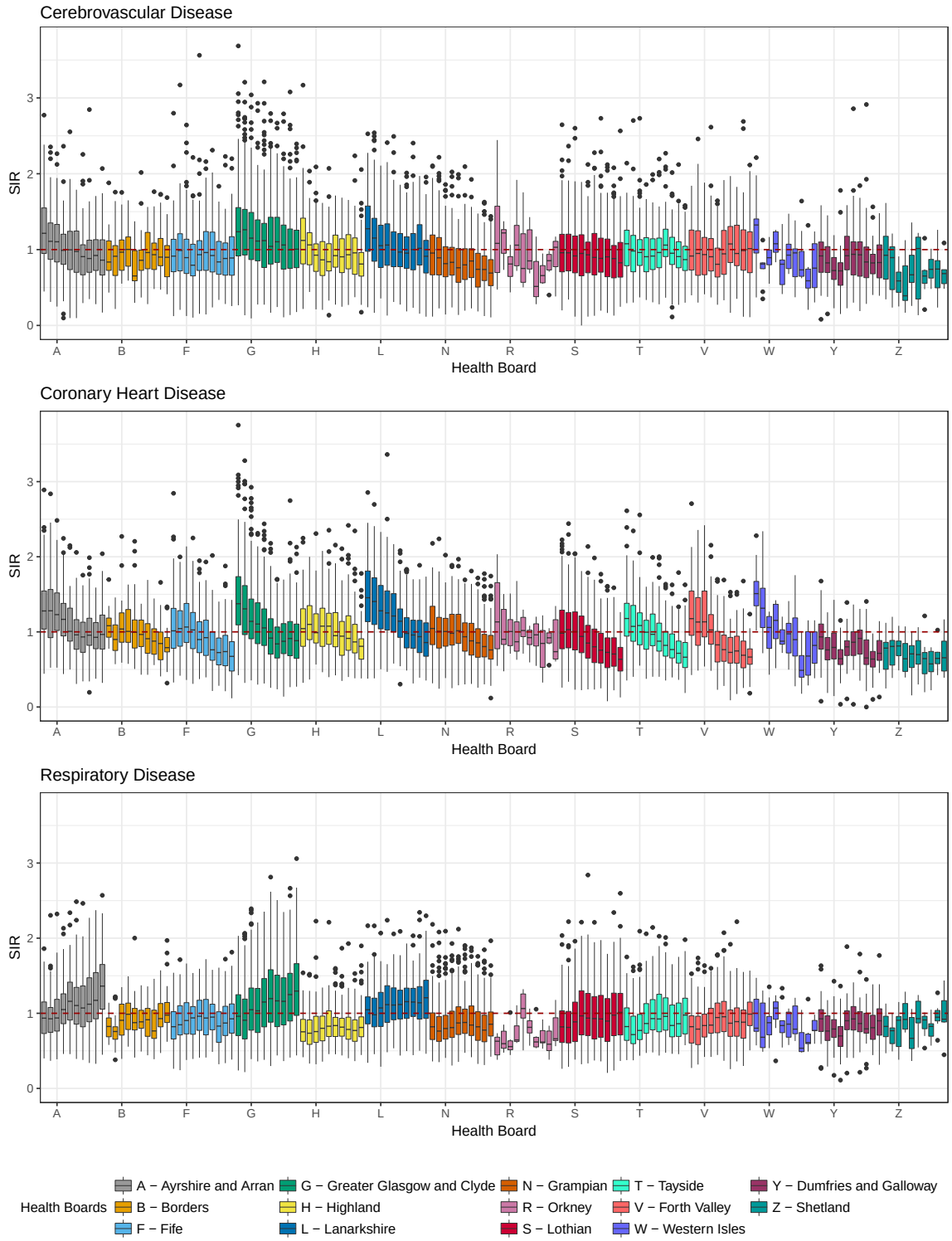

Figure 4.2: Boxplots of Standardised Incidence Ratios (SIR) for cerebrovascular disease, coronary heart disease and respiratory disease for IGs in each health board at each year (2003-2012).

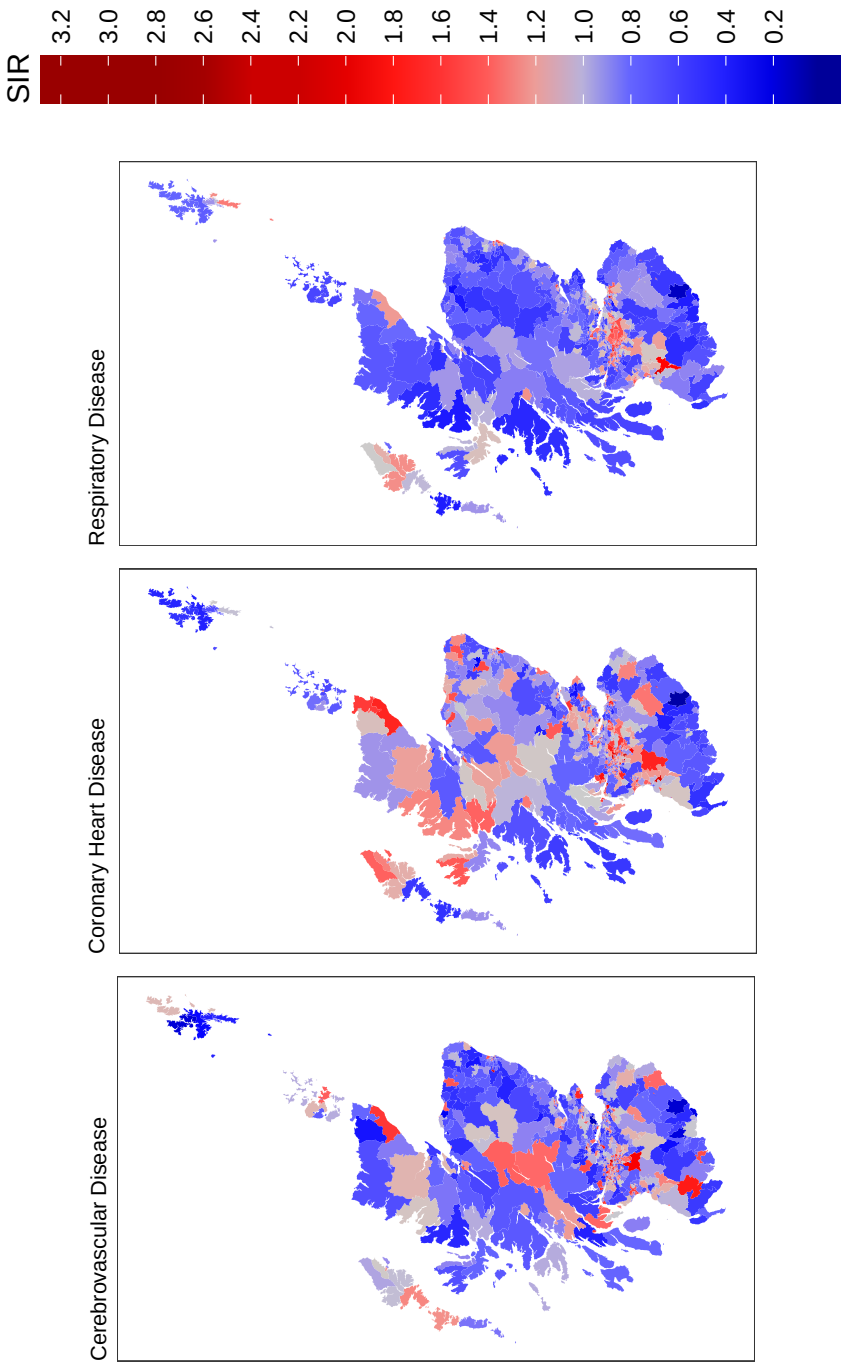

Figure 4.3: Standardised Incidence Ratios (SIR) for cerebrovascular disease, coronary heart disease and respiratory disease for each IG in Scotland in 2006.

coronary heart disease and respiratory disease. From these maps we can see that the spatial patterns for each disease are not the same. Both cerebrovascular disease and coronary heart disease have more areas in northern Scotland with high disease risk than respiratory disease. This suggests having disease specific spatial risk surfaces in the modelling.

Due to the large number of IGs located in central Scotland, it is difficult to see any pattern for this part of Scotland in these maps. Figure 4.4 shows separate maps for HBs Greater Glasgow and Clyde, Lothian and Lanarkshire, which are all located in central Scotland, to give a clearer picture of spatial patterns in these areas. One similarity between these three maps is that there are more areas in the west of Scotland which show SIR levels of above 1 compared to the east, particularly in the city of Glasgow. As seen in the maps for all of Scotland, the spatial patterns for each disease are not the same, again suggesting disease specific spatial risk surfaces.

In this study, a multivariate approach has been deemed suitable given the difficulty in measuring a population's health. It is therefore assumed that using a combination of data from three of Scotland's biggest killers would provide a better understanding of health inequalities in Scotland rather than just concentrating on a single disease. The hospital admissions for these three diseases are therefore our dependent variables. Figure 4.5 shows scatterplots of the three diseases plotted against one another to give a better understanding of the relationship between these diseases. From these plots, and the correlations which are printed on the top-right corner of each graph, we can see that for all combinations of the three diseases there are moderate positive relationships between them with correlations of around 0.5. These correlations are assumed to be caused by common factors which further justifies the reasoning to use a multivariate approach to model health inequality, since all three diseases are related to one another but including each will provide different information to better estimate health inequalities in Scotland.

In order to assess the presence of residual spatial correlation in the data, a Poisson generalised linear model (GLM) was fitted to the data for 2003 for each disease separately, with the covariates described. Moran's I (Moran, 1950) statistics were then calculated using the residuals from these models, and the results show that for

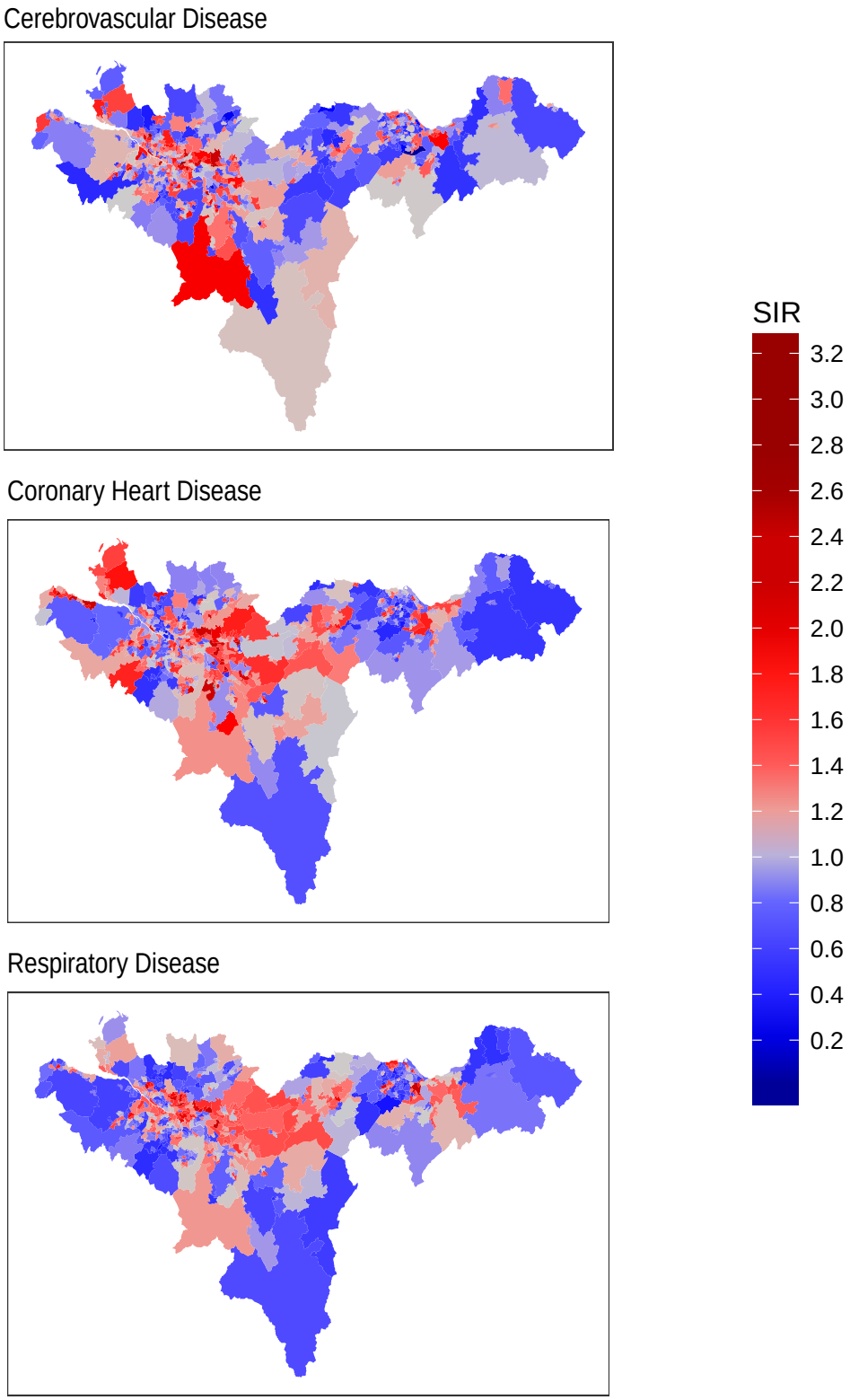

Figure 4.4: Standardised Incidence Ratios (SIR) for each IG in health boards Greater Glasgow and Clyde, Lothian and Lanarkshire in 2006 for cerebrovascular, coronary heart and respiratory disease.

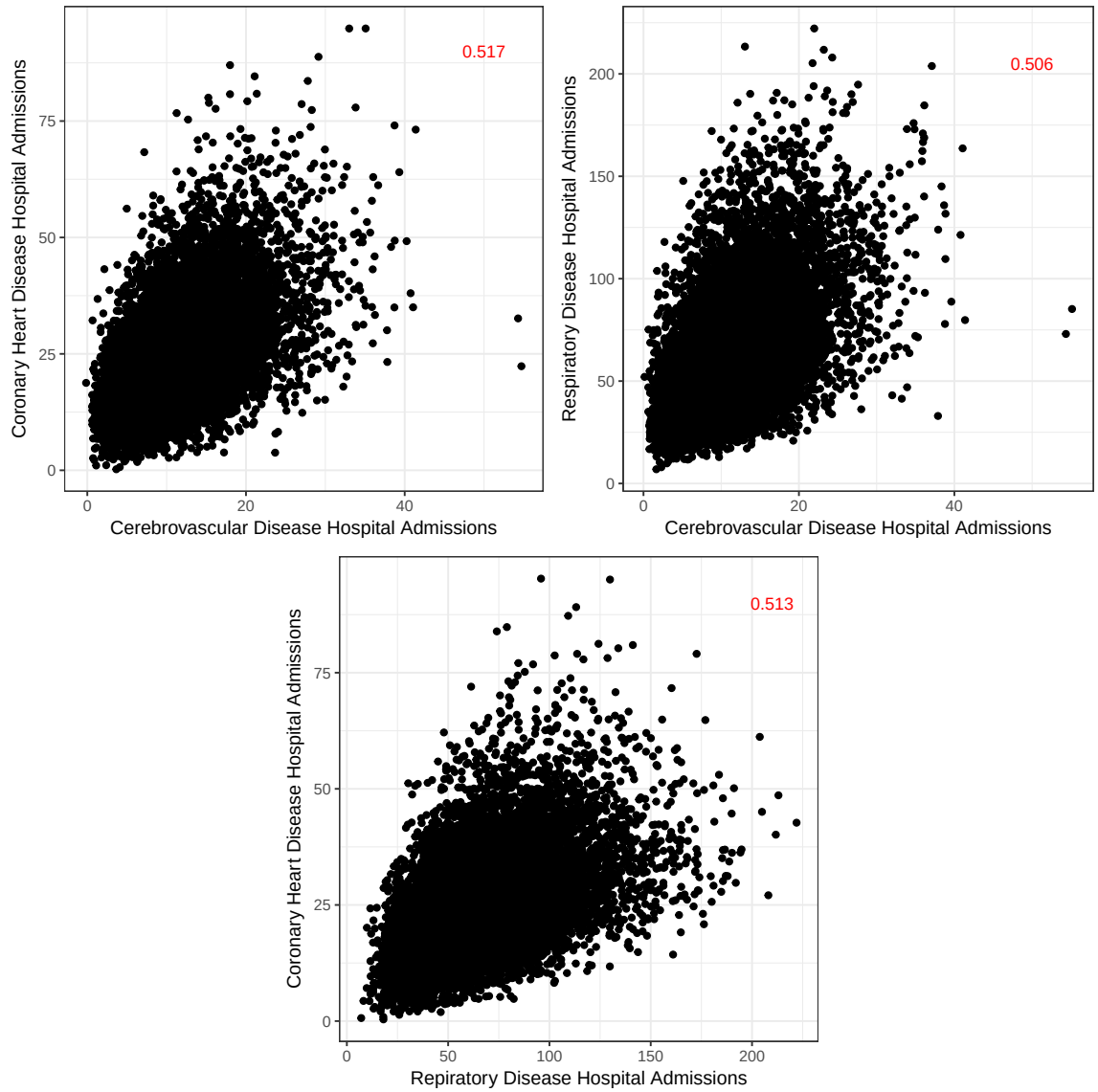

Figure 4.5: Scatterplots to show the relationship between each of the three disease. Correlations are printed in the top right of each plot.

all three diseases strong spatial correlation was present, with Moran's I statistics of 0.107, 0.227, and 0.241 for cerebrovascular, coronary heart and respiratory disease respectively, with significant associated p-values < 0.001 for all 3 statistics.

Pairwise correlations were also calculated between the residuals for each disease, with a correlation of 0.211 between cerebrovascular and respiratory disease, 0.216 between cerebrovascular and coronary heart disease and 0.337 between coronary heart and respiratory disease. This indicates that there is some residual between disease correlation in the data, suggesting between disease correlation needs to be modelled. To investigate whether disease specific covariate effects would be appropriate, the estimated covariate effects from the same Poisson GLM's as before were checked and some differences in these were found. For example the covariate log(% of population of Asian ethnicity) showed a significant protective effect for coronary heart disease, a significant increased risk effect for respiratory disease and no significant effect for cerebrovascular disease. It is therefore appropriate to include separate covariate effects for each disease.

Finally, in order to assess the presence of temporal correlation, the average lag-one correlation coefficient was calculated for each disease across the IGs. However, given that we have a very short time series (only 10 time points) the results from this were inconsistent. Given that the data come from the same group of people every year, *a priori*, we would expect there to be temporal correlation and so we will account for this in the final model.

4.3 Methodology

Here, we outline the multivariate spatio-temporal model developed to answer the questions of interest in this chapter. To extend the model used in Chapter 3 to account for multiple diseases, between disease correlation needs to be incorporated into the model. Similar to before, a hierarchical Bayesian model based on observed counts of hospital admissions, Y_{itd}, letting i denote IG ($i = 1, ..., 1235$), t denote year since 2003 ($t = 1, ..., 10$) and d denote disease ($d = 1-$cerebrovascular disease$, 2-$coronary heart disease$, 3-$respiratory disease). Expected counts, e_{itd}, for area i, time point t

and disease d were calculated using standardisation to account for differences in the demographic structures of each area.

4.3.1 Likelihood Model

The first level of the hierarchical model we specify is given by

$$Y_{itd} \sim \text{Poisson}(e_{itd}\theta_{itd}), \quad i = 1, ..., n(= 1235); t = 1, ..., T(= 10); d = 1, 2, D(= 3),$$
$$\ln(\theta_{itd}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_d + \mathcal{H}_{h(i)td} + \phi_{id}, \quad h(i) = 1, \ldots, H(= 14), \tag{4.1}$$

where Y_{itd} and e_{itd} are the observed and expected numbers of hospital admissions in IG i, time point t and disease d, while θ_{itd} is the risk relative to the expected numbers e_{itd}. We model the log-risk with 3 components, the first of which is the $p \times 1$ vector of known covariates $\boldsymbol{x}_i = (1, x_{i2}, ..., x_{ip})$, including an intercept term, with disease specific regression parameters $\boldsymbol{\beta}_d = (\beta_{1d}, ..., \beta_{pd})$. Given that we do not have access to temporally-varying covariate information we cannot include this in the model. However, we did consider allowing the regression parameters to vary over time and disease, i.e $\boldsymbol{\beta}_{td} = (\beta_{1td}, ..., \beta_{ptd})$ but the parameter estimates showed little change over time and the results can be found in Appendix B, Section B.1. The prior specified is $\boldsymbol{\beta}_d \sim N(0, 100\boldsymbol{I})$ which is weakly informative to allow their values to be informed by the data. The remaining 2 components are a baseline disease specific spatial effect ϕ_{id}, and a disease specific health board temporal trend $\mathcal{H}_{h(i)td}$, where $h(i)$ denotes that IG i is located within HB h. Both of these components are described in the following sections.

4.3.2 Disease specific spatial effects

In Section 4.2.1, we found evidence of substantial residual spatial correlations in the data, which we model via disease specific spatial random effects. Spatial correlation is induced into these random effects via the neighbourhood matrix $\boldsymbol{W}$, which is an $(n \times n)$ binary matrix, where $w_{ij} = 1$ if two areas are defined to be neighbours and $w_{ij} = 0$ if not. Also $w_{ii} = 0$ for all i. We use the same neighbourhood matrix $\boldsymbol{W}$ as in Chapter 3, i.e. the way we define if two areas are deemed to be neighbours is if

they share a common border.

Since there was evidence from Figure 3.3 that a common spatial surface may not be appropriate for all diseases, we model the multi-disease spatial effects $\boldsymbol{\phi}$ by a multivariate version of the Leroux CAR prior (Leroux et al., 2000) given by

$$\begin{aligned}
\boldsymbol{\phi}_i|\boldsymbol{\phi}_{-i} &\sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\boldsymbol{\phi}_j}{\rho\sum_{j=1}^{n} w_{ij}+(1-\rho)}, \frac{1}{\rho\sum_{j=1}^{n} w_{ij}+(1-\rho)}\boldsymbol{\Sigma}\right), \qquad (4.2)\\
\boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}(3, \boldsymbol{I}),\\
\rho &\sim \text{Unif}(0,1),
\end{aligned}$$

where $\boldsymbol{\phi}_i = (\phi_{i,1}, ..., \phi_{i,D})$ and $\boldsymbol{\phi}_{-i} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_{i-1}, \boldsymbol{\phi}_{i+1}, ..., \boldsymbol{\phi}_n)$. The parameter ρ controls the level of spatial correlation in the data, with $\rho = 0$ corresponding to independence in space and $\rho = 1$ corresponding to the multivariate extension of the intrinsic CAR prior (Besag et al., 1991). Disease-specific spatial correlation, $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$, was considered but analyses on each disease separately suggested a single ρ parameter was sufficient. The covariance matrix, $\boldsymbol{\Sigma}$, is included to allow for between disease correlation. Given there is no particular reason for an *a priori* structure for this matrix, an unconstrained form is assumed. The conjugate inverse-Wishart prior is assigned to $\boldsymbol{\Sigma}$ to allow this step to be implemented using Gibbs sampling, with weakly informative hyperparameters which will allow for these parameters to be estimated mainly by the data.

4.3.3 Temporally varying HB effects

A key question in our analysis is to investigate the health inequalities between Scotland's 14 regional health boards, and how these change over time and between disease. Therefore we include disease specific health board temporal trends in the model, $\boldsymbol{\mathcal{H}}_{hd} = (\mathcal{H}_{h1d}, ..., \mathcal{H}_{hTd})$, which are modelled by the first-order autoregressive process

$$\begin{aligned}\mathcal{H}_{htd} &\sim N(\alpha_d \mathcal{H}_{h,t-1,d}, \sigma_d^2), \\ \sigma_d^2 &\sim \text{Inverse-Gamma}(0.001, 0.001), \\ \alpha_d &\sim \text{Unif}(0, 1),\end{aligned} \tag{4.3}$$

where $\mathcal{H}_{htd}$ is the effect for health board h at time point t for disease d. Temporal correlation is induced via the hyperparameter α_d, with $\alpha_d = 0$ indicating independence across time while $\alpha_d = 1$ indicates strong temporal dependence. A previous version of this model allowed the hyperparameters α_d and σ_d^2 to vary by disease and health board within disease, however in this case the parameters were not well identified by the data, and so a simpler prior with α_d and σ_d^2 only varying by disease was implemented instead. As before, weakly informative priors were assigned to α_d and σ_d^2 to allow their values to be mainly informed from the data, and both steps to be updated using Gibbs sampling.

4.4 Estimation

Similarly to in Chapter 3, samples were drawn from the posterior distribution using Markov chain Monte-Carlo (McMC) simulation using both Gibbs sampling and Metropolis steps. The McMC algorithm was written (as part of this thesis) in `R` (R Core Team, 2014) and the `R` package `Rcpp` was again utilised to allow some of the more computationally intensive steps to be written in the more efficient language, `c++` (Eddelbuettel and François, 2011, Eddelbuettel, 2013). To make this research reproducible the code and data are available at `https://github.com/eilidhjack/MVST-software`. The updates for the majority of these parameters are very similar to those described in Section 3.4 and so, with the exception of $\boldsymbol{\Sigma}$, little detail is given about the updates for the model developed here.

4.4.1 Update for $\boldsymbol{\beta}_d$

A Metropolis step is used to sample $\boldsymbol{\beta}_d$ similar to that shown in Section 3.4.1. $\boldsymbol{\beta}_d = (\beta_{1d}, ..., \beta_{pd})$ is drawn as a single block for all 4 covariates for each disease seperately. Each of the continuous covariates were mean centered before being added to the model to allow for easier interpretation of the HB effects.

4.4.2 Update for $\boldsymbol{\phi}_i$

Each $\boldsymbol{\phi}_i$ is drawn separately using a Metropolis step similar to that shown in Section 3.4.2. Due to indentifiability issues, each ϕ_{id} was mean centered by health board and disease.

4.4.3 Update for $\boldsymbol{\Sigma}$

The between disease covariance matrix, $\boldsymbol{\Sigma}$, is drawn using Gibbs sampling as follows:

$$\begin{aligned} f(\boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \mathcal{H}, \boldsymbol{\phi}) &\propto N(\boldsymbol{0}, [\boldsymbol{Q}(\rho, \boldsymbol{W}) \otimes \boldsymbol{\Sigma}^{-1}]^{-1}) \text{Inverse-Wishart}(a, \boldsymbol{I}), \\ &\propto \text{Inverse-Wishart}(\tilde{a}, \boldsymbol{B}), \end{aligned} \tag{4.4}$$

where,

$\tilde{a} = a + n,$

$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{\phi}\boldsymbol{Q}(\rho, \boldsymbol{W})\boldsymbol{\phi}^\top$

4.4.4 Update for ρ

Finally, the spatial correlation parameter, ρ, is drawn using a Metropolis step similar to that described in Section 3.4.4.

4.4.5 Update for $\mathcal{H}_{htd}$

A Metropolis step is also used to sample the vector of $\boldsymbol{\mathcal{H}}_{hd}$ effects, similar to that shown in Section 3.4.5, where $\mathcal{H}_{htd}$ is updated separately for each health board at

each time point in each disease. The HB effects were mean centered by disease due to identifiability issues.

4.4.6 Update for σ_d^2 and α_d

Similarly to the steps described in Sections 3.4.6 and 3.4.7, σ_d^2 and α_d are drawn using Gibbs sampling, separately for each disease.

4.5 Results

The multivariate spatio-temporal model proposed in Section 4.3 was applied to the data for Scotland described in Section 4.2. Inference is based on a single McMC chain with 150,000 iterations, 50,000 of which were discarded for the burn-in period. The chain was thinned by 5 due to limitations in computer memory and to make the samples closer to independent, and so the posterior estimates are based on 20,000 samples. Convergence was checked both by examining parameter trace plots and Geweke diagnostics (Geweke, 1992).

4.5.1 Correlation

The posterior medians and 95% credible intervals for the spatial and temporal correlation parameters are shown in Table 4.1. The posterior median estimate for the spatial correlation parameter, ρ, is 0.432, which suggests a moderate level of spatial correlation across Scotland for the three diseases. A separate temporal correlation parameter α_d is assigned to each disease, and the table shows similar estimates for coronary heart disease and respiratory disease, with posterior medians of 0.870 and 0.833 respectively. Although the posterior median for cerebrovascular disease isn't as high (0.689), it still shows that the data contain moderate levels of temporal correlation, indicating that disease risks change smoothly year on year.

The covariance matrix $\mathbf{\Sigma}$ represents the conditional covariance between the disease specific random effects given the random effects at the remaining areas, and thus can be used to give a measure of the correlation between the residual (after covariate adjustment) risk surfaces between the two diseases. For example, the posterior

Table 4.1: Estimates and 95% credible intervals for spatial, temporal and between disease correlations.

Spatial Correlation	**Posterior Median**	**95% CI**
ρ	0.432	(0.360, 0.512)
Temporal Correlations	**Posterior Median**	**95% CI**
α - Cerebrovascular Disease	0.689	(0.562,0.799)
α - Coronary Heart Disease	0.870	(0.790,0.946)
α - Respiratory Disease	0.833	(0.720,0.940)
Between Disease Residual Risk Surface Correlations	**Posterior Median**	**95% CI**
Cerebrovascular and Coronary Heart	0.498	(0.465, 0.525)
Cerebrovascular and Respiratory	0.559	(0.533, 0.578)
Coronary Heart and Respiratory	0.645	(0.633, 0.658)

residual correlation between cerebrovascular and coronary heart disease risk surfaces is calculated as

$$\frac{\Sigma_{12}}{\sqrt{(\Sigma_{11}\Sigma_{22})}}. \tag{4.5}$$

Table 4.1 shows the posterior medians of the between disease correlations, along with the 95% credible intervals. All pairs of diseases show moderate correlation with one another, with the correlation between coronary heart disease and respiratory disease being the strongest with a posterior median of 0.645, while the correlation between coronary heart disease and cerebrovascular disease is the weakest with a posterior median of 0.498. This justifies our use of a multivariate modelling approach which accounts for this correlation and allows for the diseases to borrow strength from each other.

4.5.2 Health board effects

In order to investigate whether there are health inequalities between Scotland's 14 regional health boards within each disease and how these are changing over time (Question 1, Section 4.1), Figure 4.6 shows the posterior medians for each health board effect on the risk scale, $\theta_{htd} = \exp(\mathcal{H}_{htd})$, for each disease separately, after adjusting for the known covariates. For all diseases it can be seen that there are health inequalities between the HBs, as there are differences between the estimated HB posterior medians within each disease. The risk of disease is not consistent between health boards, nor is it constant over time. The way these inequalities are changing over time also differs between diseases. For cerebrovascular disease we see a general decreasing trend, and after around 2007 the trends seem to level off. We also see a narrowing of the inequality between the HBs with a difference of 0.579

between the highest and lowest median HB risk in 2003 compared to 0.184 in 2012. Given that the island boards (Orkney (R), Western Isles (W) and Shetland (Z)) have significantly fewer IGs than the mainland boards, we tend to see greater variation in risk estimates for these boards over the time period. However, even when ignoring these boards, there is still a reduction in the inequality between HBs from 0.326 in 2003 to 0.184 in 2012.

For coronary heart disease a much stronger decreasing trend can be seen over almost all of the HBs compared to cerebrovascular disease. We also see a narrowing of the inequality between the HBs, with a range of medians in 2003 of 0.676 compared to 0.354 in 2012. As with cerebrovascular disease, after removing the island boards the range in inequality lessens, but is still present, from 0.461 in 2003 to 0.270 in 2012.

Finally, the HB effects for respiratory disease do not show the same pattern as the previous two diseases. In general, most HB risks seem to go up over the time period, which is consistent with what was found in the raw data in Section 4.2.1. We also see a widening in health inequalities between the HBs for this disease, with the range between medians increasing from 0.344 in 2003 to 0.440 in 2012. However, once the island boards were removed the increase is almost non-existent from 0.318 in 2003 to 0.319 in 2012, suggesting no change in health inequalities for respiratory disease over the years between the mainland HBs.

We are also interested in comparing how average HB levels and temporal trends change between diseases within a HB (Question 2, Section 4.1). To answer this, Figures 4.7 and 4.8 show the average HB levels and temporal trends between diseases for all 14 HBs. The posterior medians (solid) and 95% credible intervals (dashed) for cerebrovascular disease are shown in red, green for coronary heart disease and blue for respiratory disease. There are some HBs whose risk for each of the three diseases is reasonably similar over the 10 years. For example, for Fife, although at the start of the time period, the posterior median for coronary heart disease was slightly above the null risk of 1, by the end, all three diseases have risk estimates of less than 1. For Dumfries and Galloway, although there is more variability at the end of the time period, the risks for all three diseases are almost always below 1 and are not

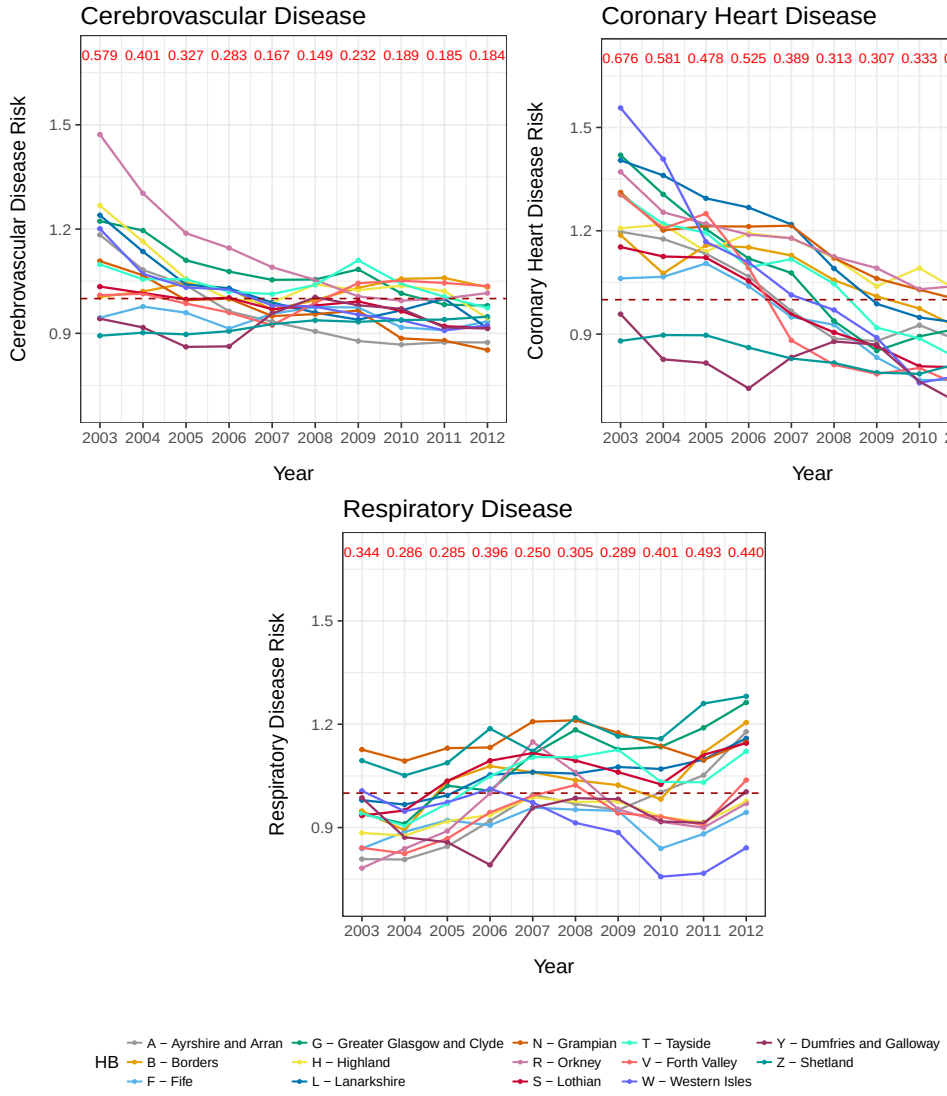

Figure 4.6: Health board risk effects across time ($\theta_{htd} = \exp(\mathcal{H}_{htd})$). Posterior medians shown for all health boards. The numbers at the top of each graph represent the range in the median HB effects for each year.

hugely different across diseases. However, from the plots it can be seen that there are differences in disease risk and patterns over time between the diseases for the rest of the HBs. For example, for Lanarkshire, a decreasing trend can be seen for coronary heart disease, and in 2003 this disease has the highest estimated risk in this HB. However, by the end of the time period the median risk for this disease is lowest in this HB. A decreasing trend can also be seen for cerebrovascular disease. However, for respiratory disease an increasing trend can be seen, and conversely to coronary heart disease the disease with the lowest risk in 2003 is respiratory disease but by the end of the time period, this disease has the highest risk for Lanarkshire. In fact, a lot of the HBs show this switch in disease risk from coronary heart disease having the highest risk in 2003 to respiratory disease having the highest risk in 2012. For Western Isles, although all three diseases show a decreasing trend, the risk for coronary heart disease was much higher than for the other two diseases in 2003, and therefore the change in risk for this disease over time is much higher than for respiratory and cerebrovascular disease. Another feature of these plots is the increased variability in the estimates for the island HBs (Orkney, Western Isles and Shetland), which can be seen by the wider credible bands and is due to them having the smallest numbers of IGs.

4.5.3 Overall health inequalities

In order to investigate whether the overall health inequalities have changed over time across the IGs in Scotland (and not just between health boards as in Section 4.5.2) (Question 3, Section 4.1), Figure 4.9 shows boxplots of the posterior median disease risk for all IGs, for each disease, from 2003-2012. Printed above each boxplot is the interquartile range (IQR) across all IGs. For cerebrovascular disease we see a general decrease in overall trend and health inequality, which can be seen from the narrowing of the boxplots and the decreasing IQRs, from 0.344 in 2003 to 0.250 in 2012. Similarly, the overall risk of coronary heart disease is decreasing over time, more quickly at the beginning of the time period and after around 2009 this decrease shows signs of leveling off. When looking at the width of the boxplots or the IQRs we can see a decrease in health inequality in coronary heart disease risk, from 0.440 in 2003 to 0.279 in 2012, which again is more noticeable in the period from 2003 to

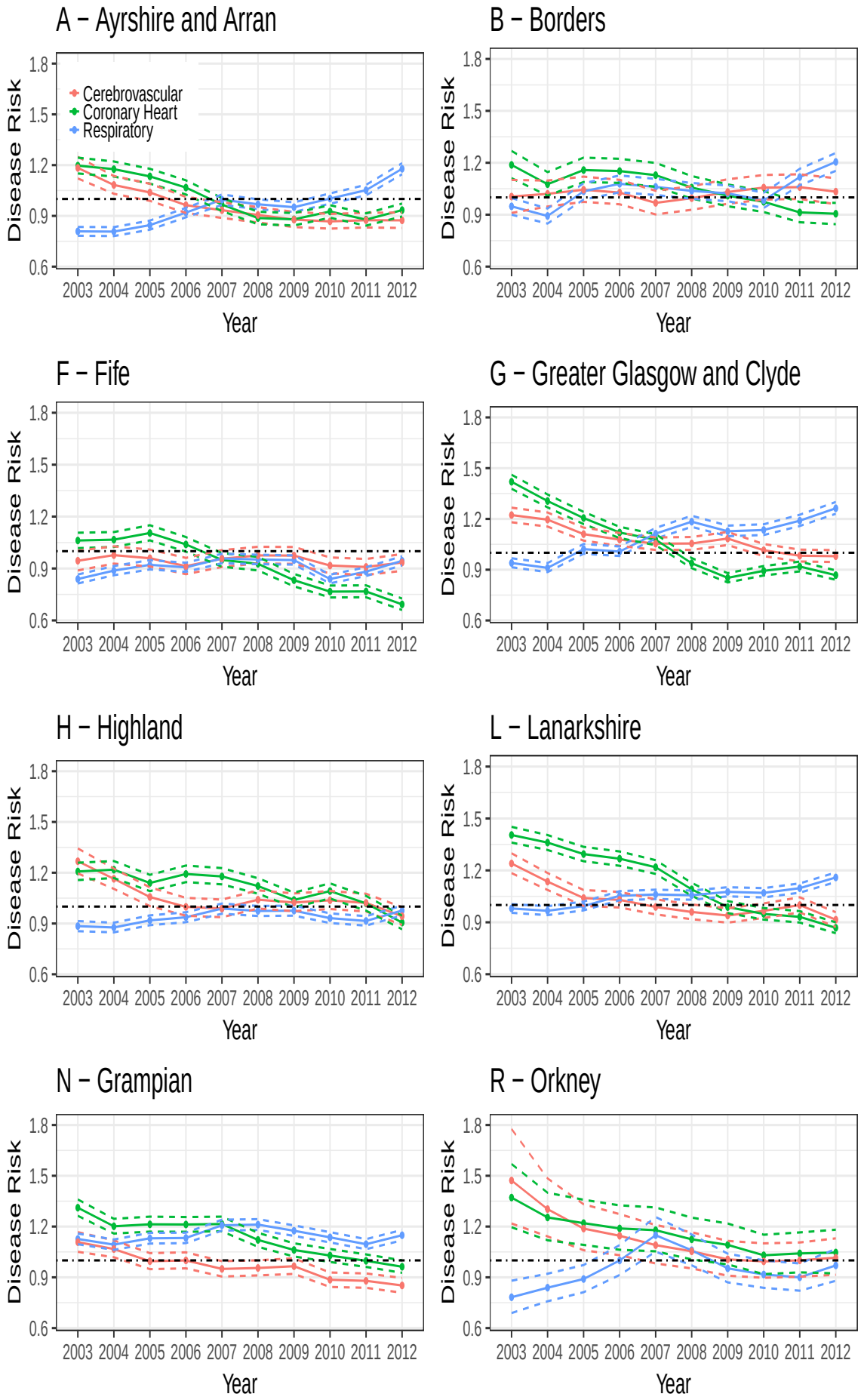

Figure 4.7: Health board risk effects across time ($\theta_{htd} = \exp(\mathcal{H}_{htd})$) for each disease. Posterior medians in red for cerebrovascular, green for coronary heart and blue for respiratory disease. Black dashed line indicates risk of 1. 95% credible intervals shown by coloured dashed lines.

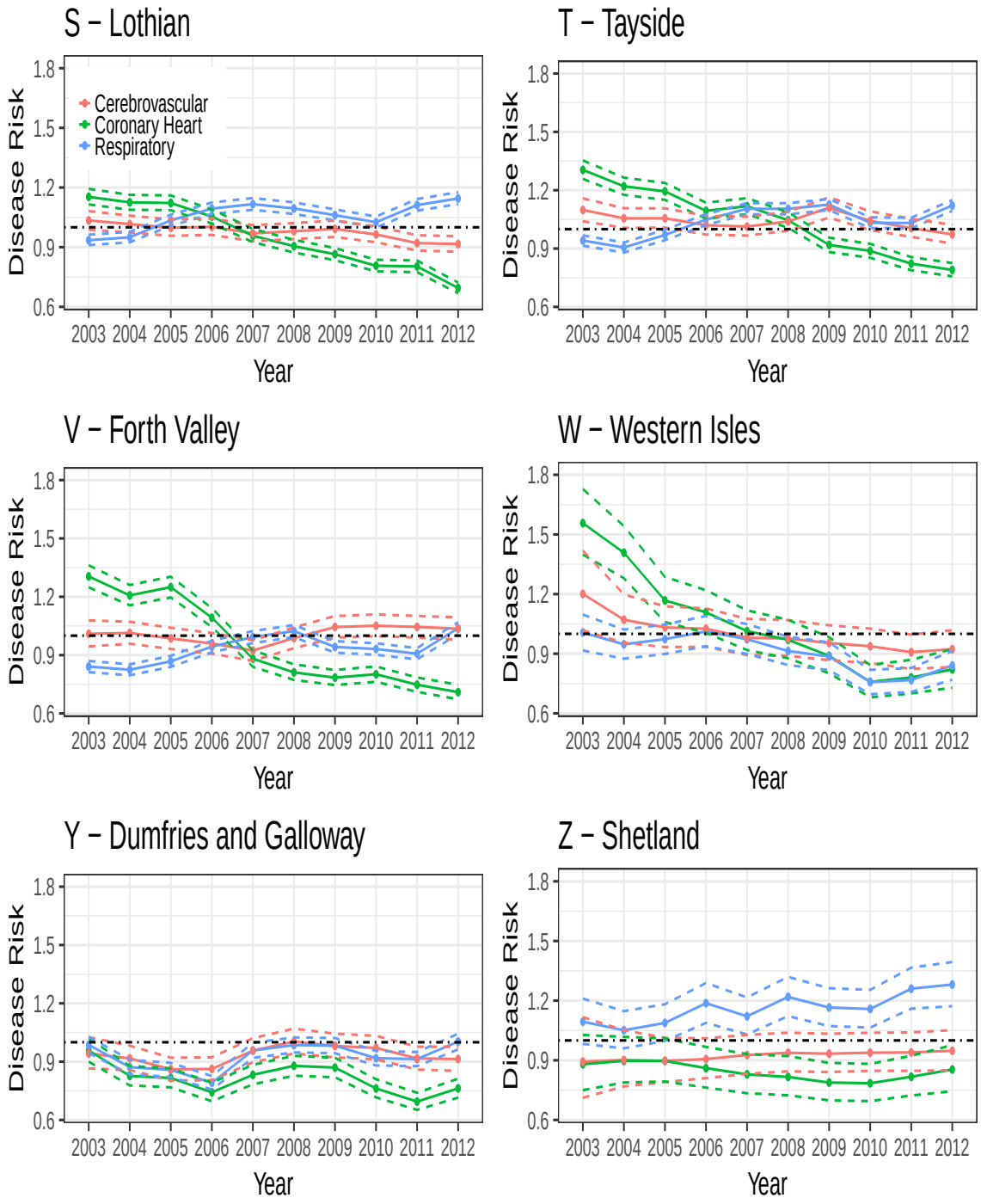

Figure 4.8: Health board risk effects across time ($\theta_{htd} = \exp(\mathcal{H}_{htd})$) for each disease. Posterior medians in red for cerebrovascular, green for coronary heart and blue for respiratory disease. Black dashed line indicates risk of 1. 95% credible intervals shown by coloured dashed lines.

2009, after which it levels off. Conversely, for respiratory disease, we see the opposite effect. Not only is the overall risk increasing over time, but the inequality is getting worse, which can be seen from the widening of the boxplots and the IQR increasing from 0.382 in 2003 to 0.532 in 2012. Some discussion on potential explanations for these results is given in Chapter 6, Section 6.4.

Figure 4.10 shows plots of the changes in uncertainty in disease risk for each of the three diseases over time. The lines in red show the proportion of IGs who have significantly higher disease risks than average (95% credible intervals that are entirely above 1), the green lines show the proportion of IGs with significantly lower disease risk than average (95% credible intervals that are entirely below 1) and the blue lines show the proportion of IGs whose 95% credible intervals contain 1 and therefore show no significant difference in risk than average. For cerebrovascular disease, the proportion of IGs with no difference in risk remains reasonably constant and the proportion of IGs in this category is higher for this disease compared to the other two. This is probably due to the fact that the hospital admissions are lower for cerebrovascular disease and therefore there is less data to estimate risk and so the uncertainty for this disease is higher. The proportion of IGs with increased risk decreases over time and the proportion with decreased risk increases over time for this disease. For coronary heart disease, the proportion of areas which show no significant difference in risk doesn't change hugely over the 10 years. The largest change is seen in the significantly higher and significantly lower risk areas, with the number of areas with significantly high risk decreasing from around 0.6 in 2003 to about 0.1 in 2012 and the number of areas with significantly low risk increasing from around 0.1 in 2003 to 0.7 in 2012. For respiratory disease, around about 0.2 of the IGs have no difference in risk across the whole period. Whereas, the number of IGs with significantly lower risk decreases from about 0.6 in 2003 to 0.4 in 2012 and the proportion of IGs with significantly higher increases. However, unlike coronary heart disease where these lines cross-over in about the middle of the time period, it isn't until 2012 when there are more areas with significantly increased risk than areas with significantly decreased risk.

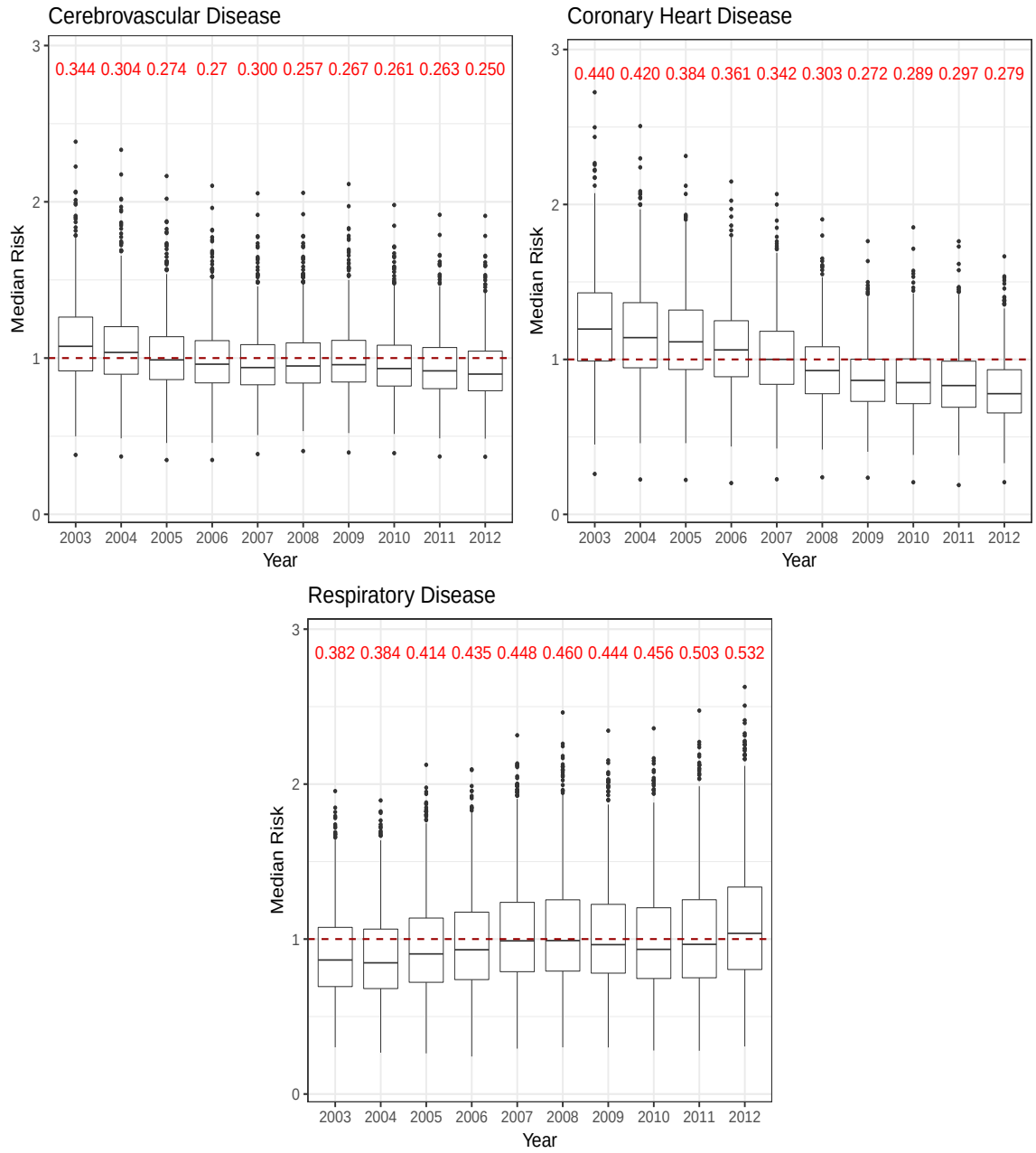

Figure 4.9: Boxplots of disease risk for cerebrovascular disease, coronary heart disease, and respiratory disease in IGs in Scotland from 2003 - 2012. The IQR across IGs are printed in red Outliers are those observations that lie outside 1.5(IQR)

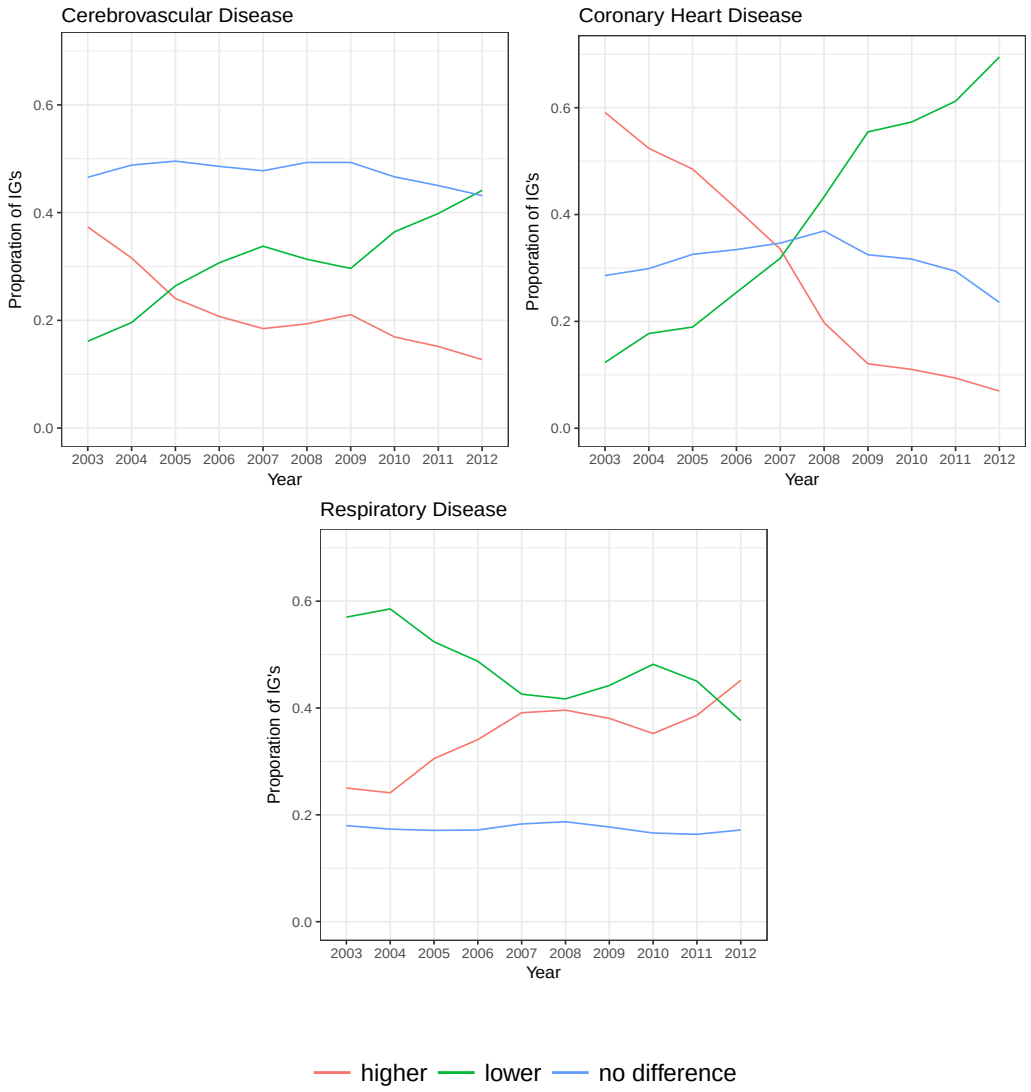

Figure 4.10: Proportion of IGs with significantly higher disease risks (95% credible interval entirely above 1) in red, proportion of IGs with significantly decreased disease risk (95% credible intervals entirely below 1) in green and proportion of IGs with no difference in risk (95% credible interval contains 1) in blue shown for each disease.

Table 4.2: Relative risk estimates for a 1% increase in each covariate (not urban/rural covariate) and 95% credible intervals for the covariates in model.

Covariate	Median RR	95% CI
Cerebrovascular Disease		
% 16-64 year olds claiming JSA	**1.060**	**(1.056, 1.065)**
Log % Asian	0.998	(0.985, 1.011)
Log % Black	**1.008**	**(1.002, 1.015)**
Rural area	0.977	(0.949, 1.006)
Coronary Heart Disease		
% 16-64 year olds claiming JSA	**1.065**	**(1.059, 1.070)**
Log % Asian	**0.965**	**(0.951, 0.980)**
Log % Black	1.001	(0.995, 1.008)
Rural area	**0.953**	**(0.924, 0.983)**
Respiratory Disease		
% 16-64 year olds claiming JSA	**1.105**	**(1.098, 1.112)**
Log % Asian	0.985	(0.968, 1.001)
Log % Black	0.992	(0.985, 1.000)
Rural area	0.997	(0.966, 1.033)

4.5.4 Covariate effects

In order to assess the impact that covariates have on risk and how this changes over disease (Question 4, Section 4.1), Table 4.2 shows the point estimates (posterior medians) and 95% credible intervals on the relative risk (RR) scale. For cerebrovascular disease the median RR for % 16-64 year olds claiming job seekers allowance is around 1.060 for a 1% increase, so the risk of cerebrovascular disease in an IG increases by 6.0% as the % claiming JSA increases by 1%. The effect of this covariate for coronary heart disease is similar to its effect for cerebrovascular disease, with a 1% increase corresponding to an increase in coronary heart disease risk of around 6.5%. However, the effect of this covariate for respiratory disease is much larger, with a 1% increase giving an increase in respiratory disease risk of around 10.5%. This could be due to the fact that smoking is one of the main causes of respiratory disease, with nearly 8 out of 10 chronic obstructive pulmonary disease (one of the most common respiratory diseases) deaths deemed to be a result of smoking (Centers for Disease Control and Prevention, 2014) and deprived areas showing higher levels of smoking compared to affluent areas (NHS Scotland, 2003a).

The covariate log(% of population of Asian ethnicity) showed no evidence of a relationship with cerebrovascular or respiratory disease risk. However for coronary

heart disease, the median RR estimate for log(% of population of Asian ethnicity) is 0.965 and the 95% credible interval is entirely less than 1, suggesting that there may be a very small decrease in coronary heart disease risk as this covariate increases. This corresponds to the results found in Chapter 3, Section 3.5.5.

Neither coronary heart disease nor respiratory disease show any evidence of a relationship between disease risk and log(% of population of Black ethnicity). However, log(% of population of Black ethnicity) was found to have a small detrimental impact of cerebrovascular disease risk, with risk increasing by 0.8% as log(% of population of Black ethnicity) increased by 1%. This is line with findings that people of black origin are at higher risk of stroke compared to white people (Stroke Association, 2016).

Finally, there is no evidence that living in a rural or urban area made any difference to the risk of cerebrovascular or respiratory disease, but for coronary heart disease the risk associated with urban areas compared with rural areas is $\frac{1}{0.953} = 1.049$, i.e. there is an estimated increased risk of coronary heart disease of 4.9% when living in an urban area compared with a rural area. Again, this is in line with the results found in Chapter 3, Section 3.5.5.

In order to assess the sensitivity of the results from this model to the choice of covariates, the model was also run with no covariates and the results in terms of risk estimates were practically identical. Some comparative figures and tables can be found in Appendix B, Section B.2.

4.5.5 Top IG risks

It was of interest to identify which IGs showed the highest risk for each disease, and hence if there were any which exhibited high risk for more than one disease. Table 4.3 shows the IGs with the top five highest risk estimates for each disease at the start of the time period (2003) and at the end (2012). Firstly, when comparing the IGs in 2003 to 2012, within disease, we notice that there are some IGs who make the top five in both these years. For example, for coronary heart disease three of the IGs with the highest risk in 2003 appear in the 2012 list, all three of which belong to the Greater Glasgow and Clyde (G) HB. For respiratory disease, two IGs remain in the top five for both years, again both of which belong to Greater Glasgow and Clyde.

Finally, all five IGs for cerebrovascular disease remain the same over the time period and all of these top five IGs belong to Greater Glasgow and Clyde. This tells us that the IGs which are at most risk of these diseases remains reasonably consistent over the time period.

When we compare the IGs across the diseases we also notice some similarities. For example, IG Paisley Ferguslie which is in Renfrewshire not only appears in the top five highest risks for all diseases in both years, but actually comes out on top for coronary heart disease and respiratory disease and is second highest for cerebrovascular disease in both years. IG Easterhouse South (full name: North Barlanark and Easterhouse South) in Greater Glasgow and Clyde appears in the top five for coronary heart disease and cerebrovascular disease for both years. Finally, Drumchapel North in Greater Glasgow and Clyde appears for respiratory disease in 2012 as well as for cerebrovascular disease in both years. It should also be noticed that 24 of the 30 IGs that appear in Table 4.3 are IGs that belong to the health board Greater Glasgow and Clyde. This highlights the extent of the deprivation, which leads to this elevated disease risk, still experienced in some areas of the Greater Glasgow and Clyde HB. These areas which have extremely high risk for two or all three diseases highlight the extent of the health inequality experienced in these areas. Ultimately this means that for the people who live in these places, the risk of hospitilisation from any one of these three diseases is much higher than average.

4.5.6 Model comparison

In order to compare our model to an existing model in the literature we decided to fit the model proposed by Quick et al. (2017b) to our data. Similar to our model this model was designed for multivariate spatio-temporal data, however unlike our model where the spatial component and temporal component are built separately, this model allows for spatio-temporal dependency in the data using a single set of random effects. The model is as follows:

Table 4.3: Posterior medians and 95% credible intervals for the top 5 IGs with the highest risk for the years 2003 and 2012 for each disease. The IGs which appear for more than one disease appear in colour.

IG Name	HB	Median Risk	95% CI
Cerebrovascular Disease (2003)			
Easterhouse South	G	2.385	(2.092,2.717)
Paisley Ferguslie	G	2.226	(1.938,2.544)
Parkhead West	G	2.064	(1.864,2.291)
Drumchapel North	G	2.060	(1.755,2.410)
Parkhead North	G	2.011	(1.810,2.231)
Cerebrovascular Disease (2012)			
Easterhouse South	G	1.910	(1.677,2.178)
Paisley Ferguslie	G	1.781	(1.554,2.043)
Parkhead West	G	1.654	(1.492,1.837)
Drumchapel North	G	1.650	(1.405,1.934)
Parkhead North	G	1.612	(1.451,1.789)
Coronary Heart Disease (2003)			
Paisley Ferguslie	G	2.724	(2.454,3.024)
Lower Bow & Larkfield	G	2.498	(2.296,2.709)
Easterhouse South	G	2.435	(2.181,2.713)
Garthamlock	G	2.267	(2.025,2.521)
Braeside	G	2.256	(2.058,2.471)
Coronary Heart Disease (2012)			
Paisley Ferguslie	G	1.665	(1.497,1.849)
Inverness Merkinch	H	1.536	(1.355,1.732)
Lower Bow & Larkfield	G	1.526	(1.403,1.659)
Braehead	A	1.510	(1.371,1.665)
Easterhouse South	G	1.489	(1.331,1.661)
Respiratory Disease (2003)			
Paisley Ferguslie	G	1.955	(1.843,2.073)
Viewpark	L	1.848	(1.751,1.950)
Heathryfold	N	1.820	(1.713,1.930)
Drumry East	G	1.794	(1.677,1.918)
Greendykes	S	1.787	(1.667,1.913)
Respiratory Disease (2012)			
Paisley Ferguslie	G	2.627	(2.478,2.786)
Doon Valley South	A	2.507	(2.357,2.665)
Drumry East	G	2.412	(2.254,2.578)
Drumry West	G	2.394	(2.246,2.550)
Drumchapel North	G	2.327	(2.171,2.494)

$$Y_{itd} \sim \text{Poisson}(e_{itd}\theta_{itd}),$$
$$\ln(\theta_{itd}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_d + \mathcal{Z}_{itd} + \phi_{itd}, \tag{4.6}$$

where $\mathcal{Z}_{itd}$ is a spatio-temporal random effect which also accounts for between disease correlation, and $\phi_{itd} \sim \text{N}(0, \tau_d^2)$. More detail can be found in Chapter 2, Section 2.8.2. The results from this model were broadly similar to ours and can be found in Appendix B, Section B.3. Due to the added complexity of the model proposed by Quick et al. (2017b) and therefore the large number of extra parameters (p_d of 11,294 compared to 3,045 for our model) the computational time was far greater than for ours. Given one of our main aims was quantifying health board inequalities, our model was more appropriate in this context.

4.6 Discussion

In this chapter a multivariate spatio-temporal model was proposed to estimate health inequalities in Scotland and how they have changed over time. The model included separate covariate effects for each disease, disease specific spatial effects and disease and health board specific temporal trends. The model was then applied to yearly hospital admissions data at the intermediate geography level for three of Scotland's biggest killers, cerebrovascular disease, coronary heart disease and respiratory disease for the period of 2003 - 2012.

The main results of this study are that, overall there has been a decrease in risk for cerebrovascular and coronary heart disease across the HBs, but this is accompanied by an increase for respiratory disease. We also found that even after the covariate effects have been removed, there still exist inequalities in disease risk between the health boards for all three diseases. These inequalities change over time and, overall they appear to be narrowing for cerebrovascular disease and coronary heart disease with a reduction in the range of the median HB risks, ignoring the island HBs, between the first and last time points of 0.142 and 0.191 for each disease respectively. However,

these inequalities show no change for respiratory disease with the difference between ranges at the start and end of the time period being only 0.001.

Overall, across the IGs in Scotland we found that health inequalities still exist to quite a considerable extent, and although there has been a narrowing for cerebrovascular and coronary heart disease, the inequalities in respiratory disease appear to be getting worse over the time period studied here. Clearly, the increase in risk of respiratory disease observed is not occurring uniformly across all IGs. Instead, risk is increasing at a higher rate for areas which were already exhibiting elevated levels of risk in 2003, which is driving the increase in the health inequality for this disease. For example, in 2003 the estimated risk for Langholm and Canonbie (lowest risk IG) was 0.302 which increased to 0.307 in 2012. In comparison, the estimated risk in 2003 for Paisley Ferguslie (highest risk IG) was 1.955 which increased to 2.627 in 2012. Therefore, in both the temporal HB trends and the overall risks estimates, the results for cerebrovascular disease and coronary heart disease show improvements both in the average risk of disease and in the health inequalities within disease, since we have shown both to be decreasing. However, the results are not as positive for respiratory disease, with increased risk over time, no change in inequality between the HB effects and an increase in inequality across all of the IGs in Scotland.

A concerning feature of our results was the large number of outliers with high risk estimates in Figure 4.9, and that there were some IGs who showed extremely high risks of disease for more than one disease as shown in Table 4.3. This further highlights the huge problem that Scotland faces in their inequality in overall health and that more needs to be done to target areas which are experiencing much higher risks of disease than the rest of Scotland.

A common problem in areal unit data of this type is that often there are changes to boundaries which occur during the time period for which data are available. For example, in 2014 The Scottish Government released a redrawn version of the intermediate geography boundaries and there are several data sets publicly available for which the time period overlaps this boundary change. Using data from before and after this change would lead to incomparable inference due to spatial misalignment in the data which would have to be dealt with. Therefore, in Chapter 5 we will

overcome this issue by utilising a common latent spatial grid scale and use a multiple imputation approach to estimate the data on this scale. Another area for future work could be to consider a clustering-based modelling approach to identify areas exhibiting elevated disease risks and investigate if these change over time and if there are common high risk clusters across the three diseases.

Chapter 5

Spatio-temporal modelling of respiratory disease risk with changing spatial boundaries

5.1 Introduction

It is not uncommon for the boundaries associated with areal unit data to change over the time period for which data are available. There are numerous reasons why such changes may take place, most commonly due to population change over time which reduces the usefulness of the original areal units. After the 2011 population census, the Scottish Government decided to redraw the boundaries of the data zones which are the key geography for small area statistics in Scotland, and are used to create the intermediate geographies which have been used so far in this thesis. The redrawn data zones were released in 2014 along with new boundaries for intermediate geographies. A map showing the differences in the boundaries between the old IGs (2001) and the new IGs (2011) is shown in Figure 5.1. Statistically, this poses a challenge since using data from before and after this change would lead to non-comparable inference due to spatial misalignment of the IG data. This chapter aims to address this problem by using a multiple imputation approach to undertake inference on a common grid for both sets of IGs, thus producing comparable inference over time. Here we adopt

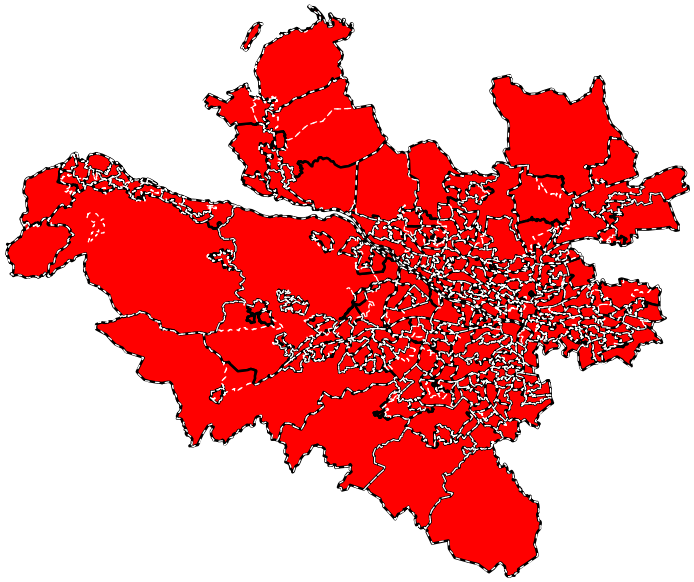

Figure 5.1: Map of 2001 IG boundaries (black) and 2011 IG boundaries (white dashed).

a common regular grid for inference, so that: (a) the results are comparable across both sets of IG boundaries over time; and (b) the results attempt to overcome the modifiable areal unit problem and each grid square has the same sized spatial support. This regular grid approach has been used by Li et al. (2012a) for these reasons, and is thus the approach I adopt here. An alternative would have been an adaptive grid, where grid square sizes varied according to population density, with larger squares in less populated regions. However, this violates point (b) above, and there are also more choices to be made, as one has to choose which size each square has, rather than just a common grid square size. A further alternative would have been to model data on the intersections between both sets of boundaries, but this again violates point (b) above and would lead to some areas being exceedingly small as the IG areas in places are almost identical. The regular grid approach will then be applied to data containing hospital admissions for respiratory disease for the years 2006 - 2016 for the health board Greater Glasgow and Clyde, where the data from 2013-2016 are reported on the redrawn IGs. We then aim to answer the following questions:

1. How has the risk of respiratory disease changed in Greater Glasgow and Clyde

from 2006 -2016?

2. How are health inequalities changing over time in Greater Glasgow and Clyde for respiratory disease risk?

We will present the results from our study and answer these questions of interest in Section 5.5. However, first the data and details on how to compute the grid level expected values are presented in Section 5.2. Section 5.3 gives detail on both of the proposed multiple imputation approaches as well as the modelling approach applied to the imputed grid data. Section 5.4 provides detail on a simulation study conducted to determine how each multiple imputation approach performs. Finally, Section 5.6 provides a discussion on the conclusions drawn from this study and possible ways in which it could be developed in the future.

5.2 Data

The study region is the Greater Glasgow and Clyde health board, which is the largest of the 14 health boards in Scotland and contains Scotland's largest city, Glasgow. The disease data are yearly counts of the numbers of hospital admissions for respiratory disease for the years 2006 to 2016 in each IG, where the boundaries for the years 2006 - 2012 follow the 2001 IG codes and the boundaries for 2013-2016 follow the 2011 IG codes. Thus let $\mathcal{A}_{k_t}$ denote the kth IG in year t, where $k_t = 1, ..., n_t$ so that there are the same set of $n_t = 254$ IGs for $t = 1, ..., 7$, (2001) and $n_t = 257$ IGs for $t = 8, ..., 11$, (2011). The disease counts are defined as $Y_t(\mathcal{A}_{k_t})$, where $t = 1, ..., 11$ denotes the time periods and $k_t = 1, ..., n_t$ denotes IG. For each year and IG, $Y_t(\mathcal{A}_{k_t})$ are the number of admissions to non-psychiatric/non-obstetric hospitals in Scotland with a main diagnosis of respiratory disease in IG $\mathcal{A}_{k_t}$ in year t, which is defined using the International Classification of Diseases Volume 10 (ICD10) codes (J00:J99, R09.1).

As well as the disease counts, our data also contains the expected number of cases collectively for each area, $e_t(\mathcal{A}_{k_t})$, which were computed using indirect standardisation. However, given that we are interested in inference on a common grid rather than on the original IGs, these need to be computed on the grid rather than at IG level.

First the common grid which will be used to make inference on is defined. A lattice of cells, $\boldsymbol{\mathcal{H}} = \{\mathcal{H}_1, ..., \mathcal{H}_M\}$ where $M = 4807$, of size $500m \times 500m$, i.e. $0.25km^2$, which covers the entire study region was produced. To assess the sensitivity of grid cell size initial research was conducted to compare 500m and 1000m grid squares. There was very little difference in the results and so the 500m grid was chosen as it leads to less granular risk surfaces. Figure 5.2 shows the grid with the 2001 IG boundaries on the top and the 2011 IG boundaries on the bottom. From this it can be seen that although the grid remains constant across the two regions, the boundaries within the regions are not. This illustrates the importance of modelling our data on the common spatial grid rather than the original IGs to obtain temporally comparable inference.

Performing inference on the new grid scale poses several challenges which need to be addressed. Firstly, the areal units $\mathcal{A}_{k_t}$ are designed to have non-zero populations and hence the IGs located in the city centre are much smaller than the IGs which are more rurally located. However, since all the grid squares are the same size, there will be some which have a population of zero since they will be areas of fields/mountains etc. where no one lives. We therefore remove these grid squares from our modelling, since it does not make sense to estimate disease risk in areas with no people. Inference is then undertaken on $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_1, ..., \mathcal{G}_m\} \subset \mathcal{H}$, where $m < M$, defined by $\boldsymbol{\mathcal{G}} = \{\mathcal{H}_i | p(\mathcal{H}_i) > 0\}$, the set of grid squares with a non-zero population. Figure 5.3 shows a map of the 2001 IG boundaries with this subsetted grid on top.

Another issue is that along the boundary of the study region, the grid squares cover areas which do not belong to the health board Greater Glasgow and Clyde. Care then needs to be taken when assigning population values to each grid square depending on what the area outwith the study region comprises of. Here $a(\mathcal{G}_i)$ is the area of grid $\mathcal{G}_i$ and $a(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$ is the area of the intersection between grid $\mathcal{G}_i$ and areal unit $\mathcal{A}_{k_t}$. Also, $p(\mathcal{G}_i)$ is the population of grid $\mathcal{G}_i$. This was obtained for $1km^2$ grids from Reis et al. (2015) and these were converted into $0.25km^2$ gridded populations to match the common grid being used for this study. This was done by assigning 1/4 of the population from each $1km^2$ grid to each $0.25km^2$ grid within it. Finally, $p(\mathcal{G}_i|\mathcal{A}_t)$ is the new adjusted population in grid $\mathcal{G}_i$ based on our study region. We propose the

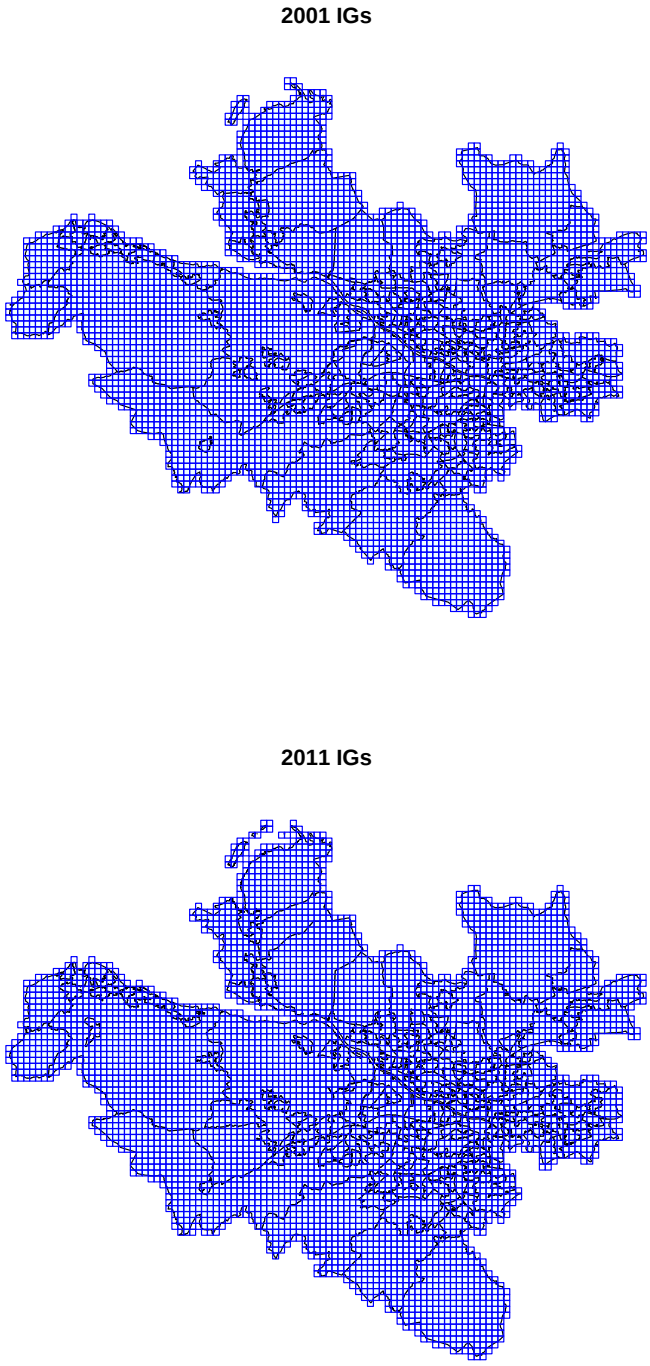

Figure 5.2: Common grid overlaid on the 2001 (top) and 2011 (bottom) IG regions.

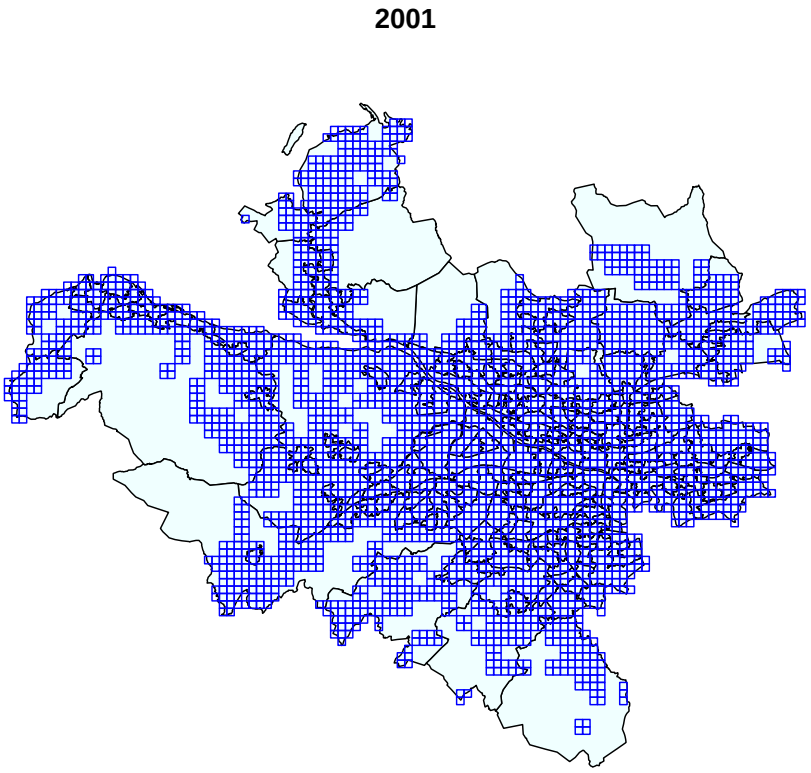

Figure 5.3: Adjusted grid overlaid on the 2001 IG regions.

following:

1. If the area in a grid square but outwith the study region is uninhabited (e.g a body of water) then we assign the entire population of that grid to our study region, i.e. $p(\mathcal{G}_i|\mathcal{A}_t) = p(\mathcal{G}_i)$.

2. If the area in a grid square but outwith the study region belongs to a bordering health board, then only a proportion of the population of that grid is assigned to the study region based on the proportion of area in the grid square that is comprised by our study region. That is $p(\mathcal{G}_i|\mathcal{A}_t) = \lfloor p(\mathcal{G}_i)\frac{\sum_{k_t=1}^{n_t} a(\mathcal{A}_{k_t}\cap\mathcal{G}_i)}{a(\mathcal{G}_i)}\rceil$, where $\frac{\sum_{k_t=1}^{n_t} a(\mathcal{A}_{k_t}\cap\mathcal{G}_i)}{a(\mathcal{G}_i)}$ is the proportion of the grid square in the study region, and $\lfloor\cdot\rceil$ denotes rounding to the nearest integer.

5.2.1 Estimating grid level expected values

To allocate the expected values at IG level, $e_t(\mathcal{A}_{k_t})$, to the common grids, $e_t(\mathcal{G}_i)$, it is clear that the total number of expected counts must be preserved, i.e. $\sum_{k_t=1}^{n_t} e_t(\mathcal{A}_{k_t}) = \sum_{i=1}^{m} e_t(\mathcal{G}_i)$. Letting $e_t(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$ be the expected number of disease counts in the intersection of areal unit $\mathcal{A}_{k_t}$ and grid $\mathcal{G}_i$, then it also follows that $e_t(\mathcal{G}_i) = \sum_{k_t=1}^{n_t} e_t(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$. Assuming then that the expected counts are distributed proportionally to the population size, we have:

$$e_t(\mathcal{A}_{k_t} \cap \mathcal{G}_i) = \frac{p(\mathcal{A}_{k_t} \cap \mathcal{G}_i)}{\sum_{j=1}^{m} p(\mathcal{A}_{k_t} \cap \mathcal{G}_j)} e_t(\mathcal{A}_{k_t}), \tag{5.1}$$

where $p(\mathcal{A}_{k_t} \cap \mathcal{G}_j)$ is the population in the intersection between grid $\mathcal{G}_i$ and areal unit $\mathcal{A}_{k_t}$ and $p(\mathcal{A}_{k_t} \cap \mathcal{G}_i)/\sum_{j=1}^{m} p(\mathcal{A}_{k_t} \cap \mathcal{G}_j)$ is the proportion of the population of area $\mathcal{A}_{k_t}$ located in grid $\mathcal{G}_i$. This quantity is unknown, however we estimate it using

$$p(\mathcal{A}_{k_t} \cap \mathcal{G}_i) = \frac{a(\mathcal{A}_{k_t} \cap \mathcal{G}_i)}{\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_i)} p(\mathcal{G}_i|\mathcal{A}_t), \tag{5.2}$$

where $a(\mathcal{A}_{k_t} \cap \mathcal{G}_i)/\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_i)$ is the proportion of geographical area in grid $\mathcal{G}_i$ that is taken up by areal unit $\mathcal{A}_{k_t}$ and $p(\mathcal{G}_i|\mathcal{A}_t)$ is the population of grid square i. This assumes that population density is constant across a grid square. For illustrative purposes, Figure 5.4 shows a simple 3×3 grid containing 4 areas. The expected counts

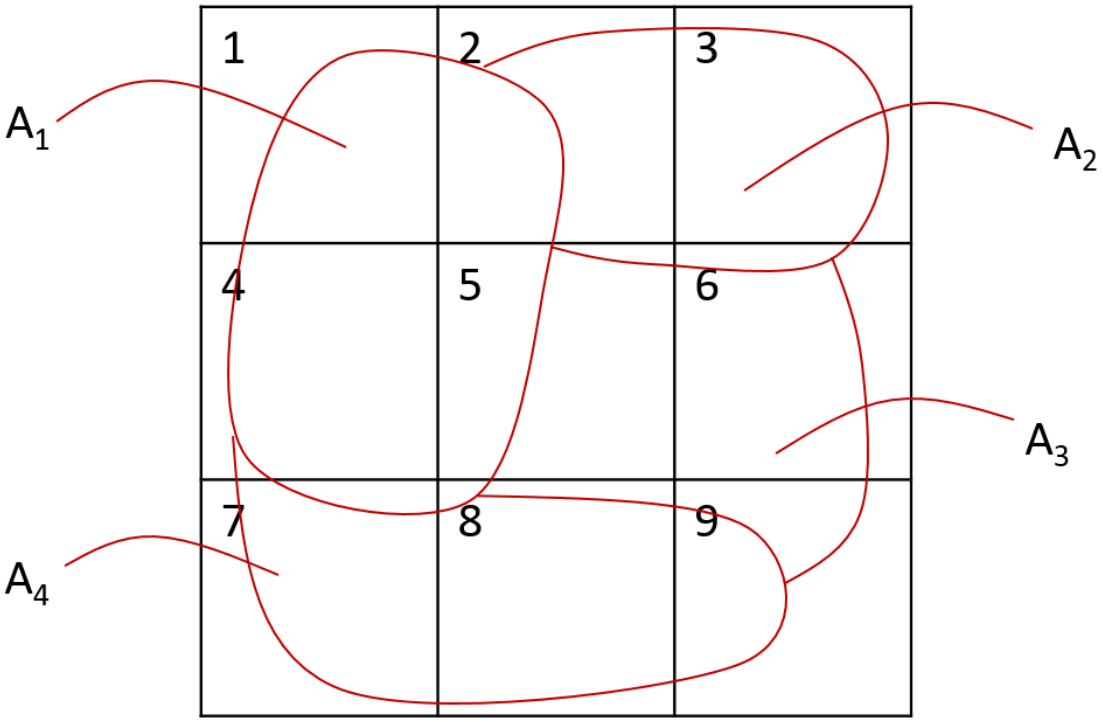

Figure 5.4: 3×3 grid containing 4 areas.

for grid number 2 is calculated as follows:

$$e_t(\mathcal{G}_2) = \frac{p(\mathcal{G}_2|\mathcal{A}_t)a(\mathcal{A}_{1_t} \cap \mathcal{G}_2)/\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_2)}{\sum_{j=1}^{m}\{p(\mathcal{G}_j|\mathcal{A}_t)a(\mathcal{A}_{1_t} \cap \mathcal{G}_j)/\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_j)\}} e_t(\mathcal{A}_{1_t}) + \frac{p(\mathcal{G}_2|\mathcal{A}_t)a(\mathcal{A}_{2_t} \cap \mathcal{G}_2)/\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_2)}{\sum_{j=1}^{m}\{p(\mathcal{G}_j|\mathcal{A}_t)a(\mathcal{A}_{2_t} \cap \mathcal{G}_j)/\sum_{r_t=1}^{n_t} a(\mathcal{A}_{rt} \cap \mathcal{G}_j)\}} e_t(\mathcal{A}_{2_t}), \tag{5.3}$$

with the remaining terms in the sum not required as their intersection areas are empty.

5.2.2 Exploratory Analysis

Figure 5.5 shows boxplots of the SIR for respiratory disease admissions in IGs in Greater Glasgow and Clyde from 2006 to 2016. Given that one of the goals of this analysis is to investigate temporal trends in respiratory disease risk, the rates for the year 2006/07 were used to calculate the expected values for all years. An increasing trend in respiratory disease risk can be seen over the time period. In 2006

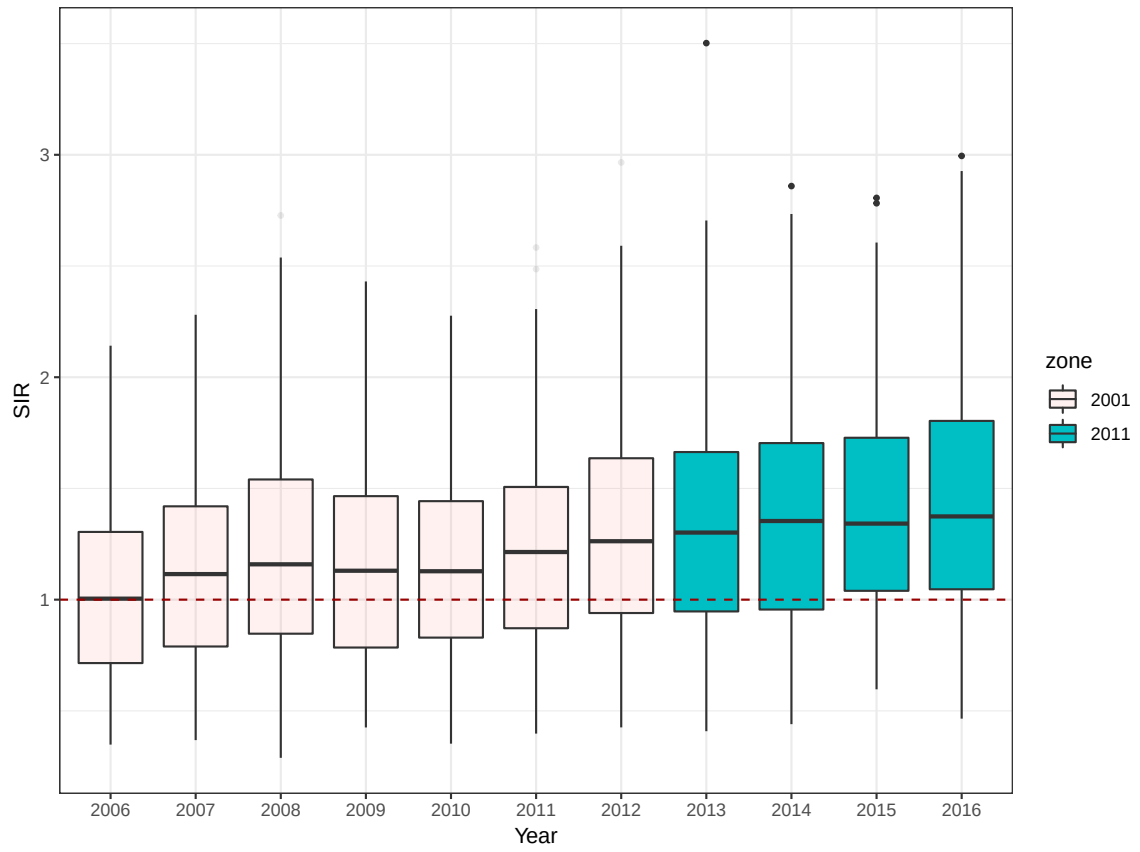

Figure 5.5: Boxplots of the standardised incidence ratio (SIR) for respiratory disease admissions for IGs in Greater Glasgow and Clyde from 2006 to 2016 by year. Years with 2001 boundaries shaded in grey. Years with 2011 boundaries shaded in green.

the median SIR is 1.005, whereas in 2016 this increases to 1.374. There also seems to be a widening in the width of the boxplots suggesting that the overall inequality in respiratory disease risk may be increasing over time. In Chapter 4 an increasing trend from 2003-2012 was found in respiratory disease, from this more recent data it is clear that the trend is respiratory disease risk continues to increase.

In order to assess the presence of spatial variation in the data, and how this has changed over the time period, Figure 5.6 shows the SIRs across IGs in Greater Glasgow and Clyde in 2006, 2010, 2013 and 2016. Again the increasing trend over time can be seen as the shading gets darker from 2006 to 2016. It can also be seen that in all 4 maps there are common areas with darker shading, indicating higher risk of respiratory disease risk, which correspond to more deprived areas of Glasgow such as the East End and Clydebank. Another notable aspect of these maps is that in areas which have higher risk of disease to begin with, it appears that the risk seems to increase at a higher rate compared to areas with low risk over the time period. This suggests that the increase in risk of respiratory disease may be more rapid for these

areas which reinforces the potential increase in health inequality which was noted in Figure 5.5.

5.3 Methodology

The following section outlines the methodology used to estimate $Y_t(\mathcal{G}_i)$ using two different multiple imputation approaches. Detail on the spatio-temporal model which is then fitted to the grid-level estimates is also outlined.

5.3.1 Multiple Imputation Approaches

We discuss two possible multiple imputation approaches for this type of data for estimating the grid level disease counts, $Y_t(\mathcal{G}_i)$, and this general approach is implemented as follows. Firstly, it is clearly true that

$$Y_t(\mathcal{G}_i) = \sum_{k_t=1}^{n_t} Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_i). \tag{5.4}$$

To estimate $Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$, the disease counts from each areal unit $Y_t(\mathcal{A}_{k_t})$ can be partitioned into the m grid square intersections $Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_1), ..., Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_m)$ using a multinomial sampling step such as:

$$[Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_1), ..., Y_t(\mathcal{A}_{k_t} \cap \mathcal{G}_m)] \sim \text{Multinomial}(n = Y_t(\mathcal{A}_{k_t}) | \omega_{k_t 1}, ..., \omega_{k_t m}). \tag{5.5}$$

Here the weights $\omega_{k_t i}$ are the probability that a disease case in area $\mathcal{A}_{k_t}$ is assigned to the intersection $(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$. The specification of the weights, $\omega_{k_t i}$, will likely depend on two quantities. Firstly, the proportion of the area of $\mathcal{A}_{k_t}$ that is comprised of the intersection $(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$, i.e.

$$\omega_{k_t i} \propto \frac{a(\mathcal{A}_{k_t} \cap \mathcal{G}_i)}{\sum_{j=1}^{m} a(\mathcal{A}_{k_t} \cap \mathcal{G}_j)}. \tag{5.6}$$

This is important since the disease counts should be allocated based on the relative size of the area of intersection between $(\mathcal{A}_{k_t} \cap \mathcal{G}_i)$ compared to the other grid squares

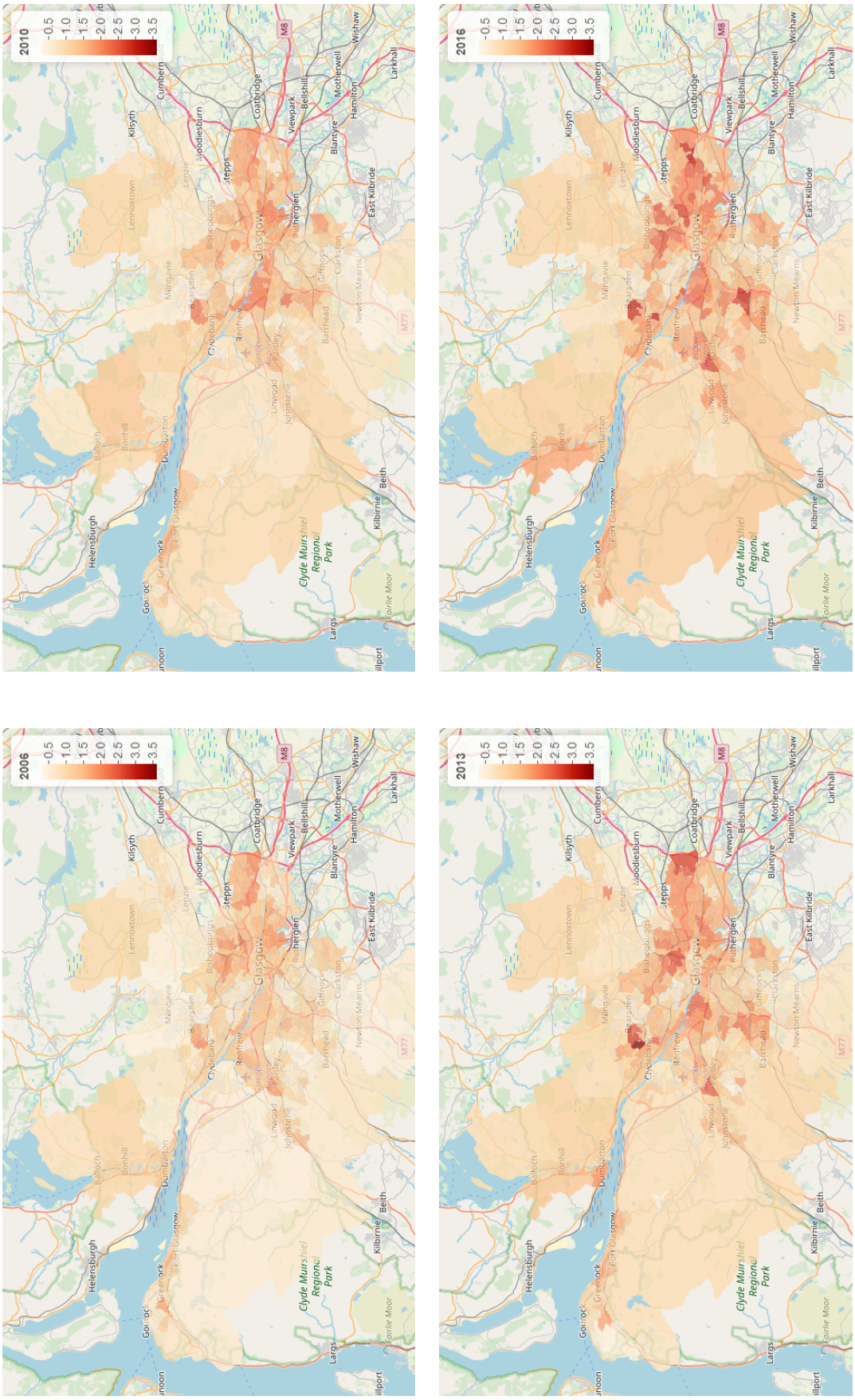

Figure 5.6: Spatial SIR maps for respiratory disease for the years 2006, 2010, 2013, 2016.

areas of intersection. Secondly, the weights should depend on the number of disease cases we would expect to observe in grid square $\mathcal{G}_i$, i.e.:

$$\omega_{k_t i} \propto e_t(\mathcal{G}_i)\theta_t(\mathcal{G}_i), \tag{5.7}$$

where $e_t(\mathcal{G}_i)$ is calculated as in Section 5.2.1. This allows the population density in each grid square to be factored in via $e_t(\mathcal{G}_i)$, as well as the estimated risk for each grid, $\theta_t(\mathcal{G}_i)$.

However, since $\theta_t(\mathcal{G}_i)$ is unknown, this quantity must be estimated before this method can be implemented. An estimate of disease risk at the areal unit level can be easily calculated using the standardised incidence ratio, $\text{SIR}_t(\mathcal{A}_{k_t}) = Y_t(\mathcal{A}_{k_t})/e_t(\mathcal{A}_{k_t})$. We therefore propose using geostatistical modelling techniques in order to predict the SIR at the grid level from the areal unit level $\text{SIR}_t(\mathcal{A}_{k_t})$, which can then be used as our grid level estimate of risk, $\theta_t(\mathcal{G}_i)$. To achieve this the spatial location of each IG, $\mathcal{A}_{k_t}$, is represented by its centroid $\boldsymbol{s}_{k_t}$ for $k_t = 1, ..., n_t$ and similarly the spatial location of each grid square, $\mathcal{G}_i$, is represented by its centroid $\boldsymbol{t}_i$ for $i = 1, ..., m$. We then use the following model, assuming that the true risk surface $\theta_t(\mathcal{G}_i)$ is spatially smooth:

$$\boldsymbol{Z} \sim \text{N}(\mu\mathbf{1}, \boldsymbol{\Sigma}(\boldsymbol{\lambda})), \tag{5.8}$$

where $\boldsymbol{Z} = (\ln(\text{SIR}(\mathcal{A}_{1_t})), ..., \ln(\text{SIR}(\mathcal{A}_{n_t})))$ is the vector of log SIR values at the n_t areal units in year t, $\mathbf{1}$ is a vector of 1s, and μ is the overall mean. The log scale is used as the SIR is non-negative and skewed to the right, and all predictions are exponentiated back to the original scale. Since spatial misalignment is the issue in the data it was decided to Krige separately each year. This avoids any additional smoothing over time when Kriging before estimating the temporal trend via the spatio-temporal modelling. The spatial autocorrelation in the data is estimated via the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\sigma^2, \tau^2, \phi)$ represent the partial sill, nugget and range parameters. The exponential covariance function is used and the model is fitted using the `geoR` package in R (Ribeiro Jr and Diggle, 2001). I chose the exponential function as it is the most commonly used one in geostatistical applications, and having

considered other options (e.g. Gaussian, spherical, etc) the estimated Kriged surfaces did not change. Within `geoR` we first computed the binned empirical semi-variogram (using the `variog()` function), which gives initial partial sill and range parameter estimates. These estimates were then used as starting points for those parameters when fitting the geostatistical model given by 5.8. We then use Kriging to predict the grid level risks, $\theta_t(\mathcal{G}_i)$, via the equations described in Chapter 2 Section 2.4. The resulting predictions, $\hat{\theta}_t(\mathcal{G}_i)$, can be used to calculate the multinomial weights using:

$$\omega_{k_t i} = \frac{e_t(\mathcal{G}_i)\hat{\theta}_t(\mathcal{G}_i)\frac{a(\mathcal{A}_{k_t}\cap\mathcal{G}_i)}{\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t}\cap\mathcal{G}_i)}}{\sum_{j=1}^{m}\{e_t(\mathcal{G}_j)\hat{\theta}_t(\mathcal{G}_j)\frac{a(\mathcal{A}_{k_t}\cap\mathcal{G}_j)}{\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t}\cap\mathcal{G}_j)}\}}, \tag{5.9}$$

which combines the two weighting elements outlined in 5.6 and 5.7. These weights were used in the multinomial imputation step described in (5.5) to estimate $Y_t(\mathcal{G}_i)$ for each grid square $\mathcal{G}_i$. However, the imputation sampling is subject to large sampling variability and to reduce this we propose drawing P realisations of the data, $\hat{Y}_t^{(p)}(\mathcal{G}) = (\hat{Y}_t^{(p)}(\mathcal{G}_1), ..., \hat{Y}_t^{(p)}(\mathcal{G}_m))$ for $p = 1, ..., P$, and then comparing two approaches of combining these data sets. Note that although this method technically uses the data twice, once to estimate $\hat{\theta}_t(\mathcal{G}_i)$ via Kriging (5.8) and then again in the multinomial imputation step (5.5), these data $Y_t(\mathcal{A}_k)$ are never used in the modelling. Instead the estimated grid level data $Y_t(\mathcal{G}_i)$ are modelled and hence it could be viewed as using the areal level data in one operation to allow for the grid level data to be estimated.

5.3.2 Approach 1: Data averaging

In the first approach the mean is taken over the P realisations as follows:

$$\hat{Y}_t(\mathcal{A}_{k_t}\cap\mathcal{G}_i) = \lfloor\frac{1}{P}\sum_{j=1}^{P}\{\hat{Y}_t^{(j)}(\mathcal{A}_{k_t}\cap\mathcal{G}_i)\}\rceil, \tag{5.10}$$

where rounding to the nearest integer is undertaken if required. With this approach the overall number of disease cases is not maintained, i.e. $\sum_{t=1}^{T}\sum_{i=1}^{m}\hat{Y}_t(\mathcal{G}_i) \neq \sum_{t=1}^{T}\sum_{k_t=1}^{n_t} Y_t(\mathcal{A}_{k_t})$ but the differences were not large. The mean was preferred to the median for averaging as it produced more accurate results, in terms of $\sum_{t=1}^{T}\sum_{i=1}^{m}\hat{Y}_t(\mathcal{G}_i) \approx$

$\sum_{t=1}^{T}\sum_{k_t=1}^{n_t} Y_t(\mathcal{A}_{k_t})$. This method therefore produces one final set of estimated grid level disease counts $\hat{\boldsymbol{Y}}(\mathcal{G})$ which is then used to fit a spatio-temporal model.

5.3.3 Approach 2: Posterior risk averaging

The second approach involves fitting a separate spatio-temporal model to each $\hat{\boldsymbol{Y}}^{(p)}(\mathcal{G})$ and then combining the estimates from each model by combining the samples from the posterior distributions for each parameter and using this to calculate any quantities needed when producing model results. For example, the posterior distribution for $\theta_t(\mathcal{G}_i)$ is:

$$f(\theta_t(\mathcal{G}_i)|\hat{\boldsymbol{Y}}) = \{f(\theta_t(\mathcal{G}_i)|\hat{\boldsymbol{Y}}^{(1)}), ..., f(\theta_t(\mathcal{G}_i)|\hat{\boldsymbol{Y}}^{(P)})\}. \tag{5.11}$$

This will now have P times as many samples as each individual model before the posterior distributions have been combined. Any quantities that need to be calculated can be done so from $f(\theta_t(\mathcal{G}_i)|\hat{\boldsymbol{Y}})$ in the usual way.

5.3.4 Spatio-temporal model

Since the overarching aim of this thesis is to quantify how health inequalities are changing over time, a spatio-temporal model which allows for an overall temporal trend, and separate spatial surfaces for each time period was deemed to be appropriate. This allows for a separate variance parameter at each time period which will allow for the comparison of health inequalities over time. A generalisation of the model proposed by Napier et al. (2016) was fitted and is of the form:

$$\begin{aligned} \hat{Y}_t(\mathcal{G}_i) &\sim \text{Poisson}[e_t(\mathcal{G}_i)\theta_t(\mathcal{G}_i)] \quad i = 1, ..., m, \quad t = 1, ..., T, \\ \ln[\theta_t(\mathcal{G}_i)] &= \beta_0 + \phi_t(\mathcal{G}_i) + \delta_t, \end{aligned} \tag{5.12}$$

where $\theta_t(\mathcal{G}_i)$ is the grid level risk of disease for the model. The overall temporal trend is represented by $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$ and this is augmented by a separate spatial surface at each time period, $\boldsymbol{\phi}_t(\mathcal{G}) = (\phi_t(\mathcal{G}_1), ..., \phi_t(\mathcal{G}_m))$. Spatial autocorrelation

is induced via an $m \times m$ neighbourhood matrix $\boldsymbol{W}$. Since we have removed grid squares with a population of zero we are unable to specify neighbours based on if two areas share a common border, like we have in previous chapters, since some areas will have no neighbours under this specification, which can be seen in Figure 5.3. Instead we use the k-nearest neighbours specification with k=4. The value of 4 was chosen since this would have been the number of neighbours most grids would have had if using the sharing a common border specification (assuming rook adjacency). However, one issue with this is that the resulting $\boldsymbol{W}$ matrix is not symmetric since it could be that $w_{ij} = 1$ and $w_{ji} = 0$ under this specification. We therefore force $\boldsymbol{W}$ to be symmetric by setting $w_{ij} = w_{ji} = \max\{w_{ij}, w_{ji}\}$ to overcome this. Similarly to previous chapters, we model the spatial random effects using the Leroux CAR prior (Leroux et al., 2000) given by

$$\phi_t(\mathcal{G}_i)|\boldsymbol{\phi}_t(\boldsymbol{\mathcal{G}}_{-i}) \sim N\left(\frac{\rho_S \sum_{j=1}^{m} w_{ij}\phi_t(\mathcal{G}_j)}{\rho_S \sum_{j=1}^{m} w_{ij} + (1-\rho_S)}, \frac{\tau_t^2}{\rho_S \sum_{j=1}^{m} w_{ij} + (1-\rho_S)}\right), \tag{5.13}$$

$$\tau_t^2 \sim \text{Inverse-Gamma}(1, 0.01),$$

$$\rho_S \sim \text{Unif}(0, 1),$$

where $\boldsymbol{\phi}_t(\boldsymbol{\mathcal{G}}_{-i}) = (\phi_t(\mathcal{G}_1), ..., \phi_t(\mathcal{G}_{(i-1)}), \phi_t(\mathcal{G}_{(i+1)}), ..., \phi_t(\mathcal{G}_m))$. The spatial dependence parameter ρ_S is common to all time points but the variance parameter τ_t^2 is allowed to vary over time. This is beneficial since the changes in variance, and therefore in inequality, are of direct interest here. The overall temporal trend $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$ is also given a Leroux CAR prior with a common temporal dependence parameter ρ_T and variance parameter τ_T^2:

$$\delta_t|\boldsymbol{\delta}_{-t} \sim N\left(\frac{\rho_T \sum_{j=1}^{T} d_{tj}\delta_j}{\rho_T \sum_{j=1}^{T} d_{tj} + (1-\rho_T)}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^{T} d_{tj} + (1-\rho_T)}\right), \tag{5.14}$$

$$\tau_T^2 \sim \text{Inverse-Gamma}(1, 0.01),$$

$$\rho_T \sim \text{Unif}(0, 1).$$

Similar to the neighbour matrix $\boldsymbol{W}$, $\boldsymbol{D} = (d_{tj})$ is a binary $T \times T$ temporal neighbourhood, where $d_{tj} = 1$ if $|j - t| = 1$ and $d_{tj} = 0$ otherwise.

5.3.5 Estimation

In order to obtain posterior summaries of each parameter, samples were drawn from the posterior distribution using Markov chain Monte-Carlo (McMC) simulation using both Gibbs sampling and Metropolis steps. Since the focus of this chapter was developing an approach for pseudo-continuous inference on a common grid rather than the development of a novel spatio-temporal model (as in Chapter 4), the model was fitted using the `CARBayesST` package(Lee et al., 2018) in `R` (R Core Team, 2014).

5.4 Simulation study

In order to determine which of the two imputation methods is best, a simulation study was conducted under two different scenarios.

- **Scenario 1** - increasing trend over time.
- **Scenario 2** - no trend over time.

For each scenario, 100 data sets were simulated on the common grid, which were then aggregated to areal unit level in order to be in the same form as the data in this study as follows:

$$Y(\mathcal{A}_{k_t}) = \sum_{i=1}^{m} \left\{ \frac{a(\mathcal{A}_{k_t} \cap \mathcal{G}_i)}{\sum_{r_t=1}^{n_t} a(\mathcal{A}_{r_t} \cap \mathcal{G}_i)} Y(\mathcal{G}_i) \right\}. \tag{5.15}$$

Each of the simulated aggregated data sets were then imputed using both of the approaches described above (see Sections 5.3.2 and 5.3.3). The bias, root mean square error (RMSE) and 95% coverage probabilities were then calculated for the corresponding risk estimates for each method as follows.

1. **Bias**

$$\text{Bias}[\theta_t(\mathcal{G}_i)] = \mathbb{E}[\hat{\theta}_t(\mathcal{G}_i)] - \theta_t(\mathcal{G}_i) \approx \frac{1}{100}\sum_{j=1}^{100} \hat{\theta}_t^{(j)}(\mathcal{G}_i) - \theta_t(\mathcal{G}_i). \tag{5.16}$$

2. **RMSE**

$$\text{RMSE}[\theta_t(\mathcal{G}_i)] = \sqrt{\mathbb{E}[\{\hat{\theta}_t(\mathcal{G}_i) - \theta_t(\mathcal{G}_i)\}^2]} \approx \sqrt{\frac{1}{100}\sum_{j=1}^{100} \{\hat{\theta}_t^{(j)}(\mathcal{G}_i) - \theta_t(\mathcal{G}_i)\}^2}. \tag{5.17}$$

3. **Coverage probability** - The percentage of the 95% credible intervals for $\theta_t(\mathcal{G}_i)$ which contain the true value for $\theta_t(\mathcal{G}_i)$.

Due to issues with memory it was decided to reduce the size of the data set to be simulated to 8 years rather than the full 11 years (which reduces the number of grids from 34507 to 25096), i.e. $t = 1, ..., 8$, corresponding to the years 2009-2016. For each method P, the number of data realisations drawn from Equation 5.5, was set to 20. These results are summarised below.

Firstly, the overall bias, RMSE and 95% coverage probabilities are shown in Table 5.1 averaged over both the 100 simulated data sets and over all elements (i.e. over all grid cells m and time points T) in each simulated data set. These results show that in both scenarios, over all risk estimates, the posterior risk averaging approach performs better than the data averaging approach in terms of bias, RMSE and coverage. The bias and RMSE are closer to 0 and the coverage is closer to the nominal 0.95 levels for this approach in both scenarios. The data averaging approach has relatively low coverage in both scenarios, suggesting that the 95% credible intervals are too narrow. This is probably since the data being fed into the spatio-temporal model, $\hat{Y}(\mathcal{G}_i)$, has been estimated and is therefore already incorrect before even being modelled. Hence, the estimation from the model is going to be poor (since it is based on an estimation

Table 5.1: Overall bias, RMSE and coverage for all risk estimates under both scenarios.

Scenario	Approach	Bias	RMSE	Coverage
Increasing trend	Data averaging	-0.01388	0.10516	0.767
	Posterior risk averaging	-0.00874	0.08106	0.924
No trend	Data averaging	-0.01647	0.11785	0.757
	Posterior risk averaging	-0.01057	0.09894	0.927

of the data) and therefore the coverage is low. Whereas with the posterior risk averaging approach, although each $\hat{Y}(\mathcal{G}_i)$ is still an estimate and therefore incorrect, we are now combining the results from 20 difference realisations of $\hat{Y}(\mathcal{G}_i)$, and so we are incorporating the uncertainty in the imputed values of $Y(\mathcal{G}_i)$ when estimating the model parameters. Hence, the uncertainty intervals are wider and the coverage higher.

We believe that the reason the posterior risk averaging approach may perfom better in terms of bias and RMSE may be partially due to the fact that, in the model averaging approach, the overall number of disease cases is not maintained (see Section 5.3.2 for details) and is always lower than the true number. Hence, in the grids where these disease counts are lost the risk may not be estimated accurately.

Figures 5.7, 5.8 and 5.9 show histograms of the bias, RMSE and 95% coverage probablities for each risk estimate averaged over the 100 simulated data sets for each approach. From Figure 5.7, it can be seen that in both scenarios, the histogram for the posterior risk averaging approach has a higher frequency of estimates with a bias close to 0. This approach also appears to be more symmetrical around 0 than the data averaging approach. The data averaging approach shows slightly more spread indicating that, on average, posterior risk averaging gives slightly less biased results. From Figure 5.8, again there appear to be slighty more estimates with an RMSE closer to 0 for the posterior risk averaging approach than the data averaging approach, although this is very marginal. This backs up the results from Table 5.1 that the posterior risk averaging approach is less biased and produces less varied results. Finally, from Figure 5.9 it is clear to see that the 95% coverage probabilities for each risk estimate using posterior risk averaging are, on average, higher than those for data averaging. This suggests that under posterior risk averaging, the 95% credible intervals are much closer to the correct width than they are using data

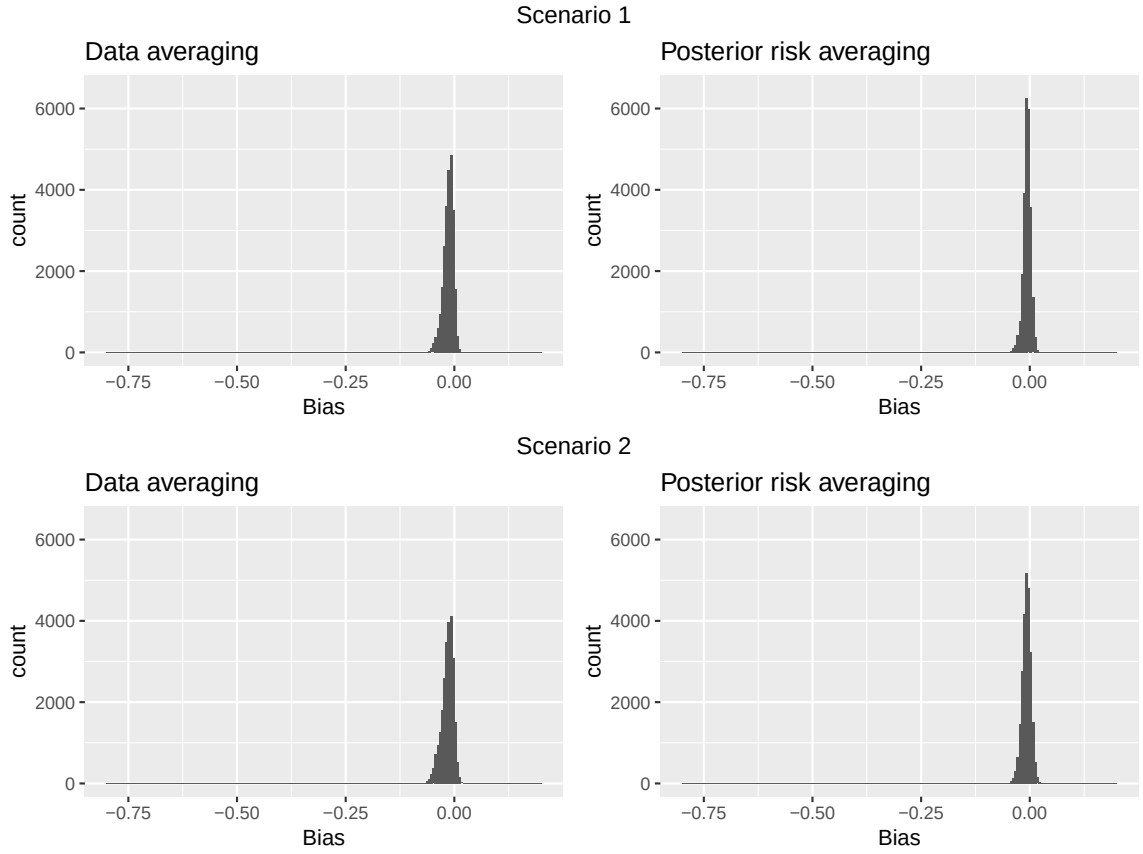

Figure 5.7: Bias for risk estimates over 100 simulated data sets for each imputation approach and each simulation scenario.

averaging. In all three figures it can be seen that in both scenarios and methods, for some estimates the modelling approach does not perform very well. This can be seen by the small number of outliers in each of the histograms. However, for the vast majority of risk estimates, on average, the posterior risk averaging approach and the following spatio-temporal modelling performs well and provides accurate results.

5.5 Results

The posterior risk averaging method proposed in Section 5.3.3 was applied to the data for Greater Glasgow and Clyde described in Section 5.2 with P, the number of data sets drawn from Equation 5.5, equal to 20. The spatio-temporal model described in Section 5.3.4, proposed by Napier et al. (2016), was then applied to each of P sets of imputed grid data. For each model a single McMC chain was run for 250,000 iterations, 150,000 of which were discarded for the burn-in period. Each chain was thinned by 50 due to limitations in computer memory and to reduce autocorrelation of the Markov chain, hence the posterior chains from each model

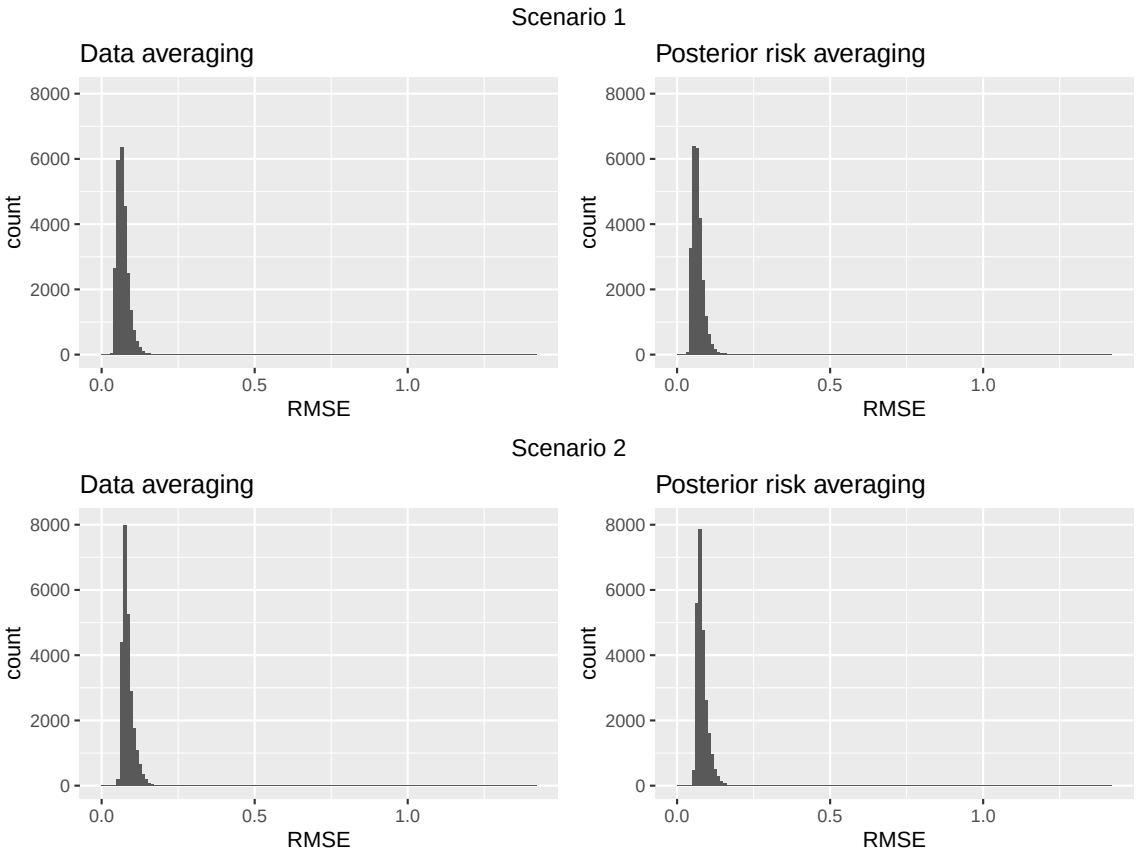

Figure 5.8: RMSE for risk estimates over 100 simulated data sets for each imputation approach and each simulation scenario.

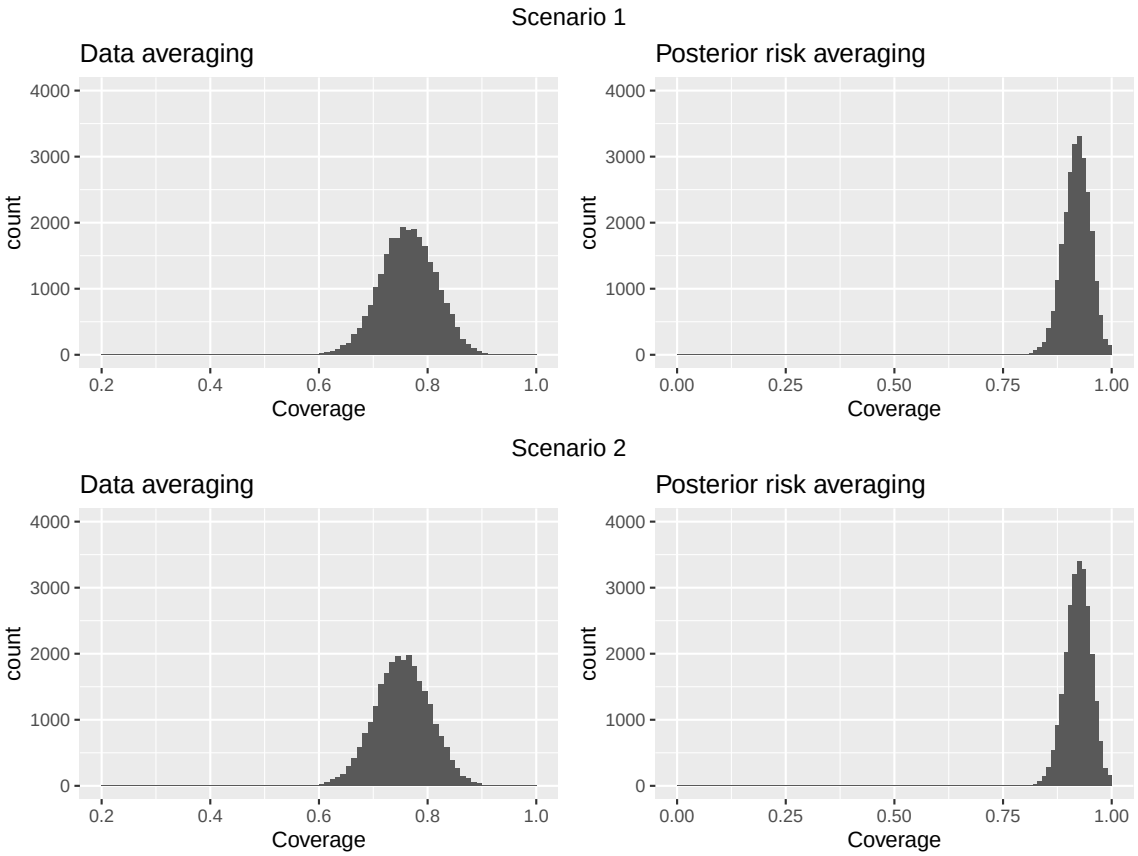

Figure 5.9: 95% coverage probabilities over 100 simulated data sets for risk estimates for each imputation approach and each simulation scenario.

contain 2000 nearly uncorrelated samples. These are then combined over the P models to calculate posterior estimates for each parameter which are based on 40,000 samples.

Of the 34507 data points on the grid scale, 15892 were estimated to be 0 (46.1%), hence the model is expected to estimate parameters with many data points providing very little information. In order to increase the information provided to the model the spatial dependence parameter, ρ_S, and the temporal dependence parameter, ρ_T, were set to be 1, thus forcing a spatio-temporally smooth surface. Strong temporal smoothness is expected because the population in a grid square is largely the same set of people each year, hence their risk of hospitalisation will be unlikely to change rapidly from one year to the next. Similarly, strong spatial correlation is expected as the grid squares are small units of $500m^2$, and hence one would expect neighbouring ones to have similar risks as a result of Tobler's first law of geography which states 'Everything is related to everything else, but near things are more related than distant things.' (Tobler, 1970). Furthermore, exploratory analysis of the data at the coarse IG level shows (via Moran's I correlation statistic) that the data contain strong spatial correlation, hence at the smaller grid level this will also occur. To assess the sensitivity of the results, the models were also run with ρ_S and ρ_T equal to 0.9 (which does not violate our belief that the data contain strong spatio-temporal correlation) and the results were very similar. Model convergence was checked both by examining parameter trace plots and Geweke diagnostics (Geweke, 1992).

5.5.1 Spatial pattern over time

In order to investigate how the risk of respiratory disease has changed in Greater Glasgow and Clyde from 2006 to 2016 (Question 1, Section 5.1), Figure 5.10 shows estimated respiratory disease risks for Greater Glasgow and Clyde for the years 2006, 2010, 2013 and 2016. Looking at each map individually clearly the risk of respiratory disease is not constant over Greater Glasgow and Clyde. Areas of high risk can be seen in each of the 4 maps which correspond to more deprived areas of the health board such as the East End of Glasgow and Clyde Bank. This is no surprise given the results from the previous Chapters where areas in Greater Glasgow and Clyde

were found to have some of the highest disease risk in Scotland. Now looking at the pattern in risk over time it can be seen that the risk of respiratory disease is increasing over the time period studied here. Although there seems to be a general increase over the whole area, it is most prominent in the areas which were already exhibiting increased risk at the start of the time period. This suggests that the rate at which the risk of respiratory disease is increasing is higher in these areas, which suggests that health inequalities are also increasing over this time frame.

When we compare these maps to the raw SIR maps shown in Figure 5.6 the maps of disease risk are much smoother. This can be explained firstly, as touched on in previous chapters, given the nature of spatial smoothing where random effects borrow strength from their neighbours, the corresponding risk estimates tend to be smoother and less extreme than the raw SIR values. However the most obvious reason for the smoothness of the risk estimates from the model is that they are now estimated on a psuedo-continuous grid rather than on the original IGs.

5.5.2 Overall health inequalities

In order to investigate how health inequalities in respiratory disease risk have changed over time in Greater Glasgow and Clyde (Question 2, Section 5.1), Figure 5.11 shows boxplots of the posterior median respiratory disease risk for all grids from 2006 to 2016. The years where data are collected on the new 2011 boundary IGs are shaded in green. Firstly from this plot we see an overall increasing trend in respiratory disease risk in Greater Glasgow and Clyde from 2006 to 2016. This is in line with what was found in Chapter 4 and indicates that the risk has continued to increase steadily from 2012 onwards.

In order to assess how health inequalities in respiratory disease risk have changed over time in Greater Glasgow and Clyde the variation in estimated disease risk is of interest. This can be assessed in Figure 5.11 by either looking at the width of the boxplots, or by examining how the interquartile range, which is printed at the top of each boxplot in red, changes over time. From this it can be seen that overall there has been an increase in health inequalitity over time in Greater Glasgow and Clyde. The IQR for 2006 is 0.357 compared to 0.444 in 2016. Not only has the IQR increased

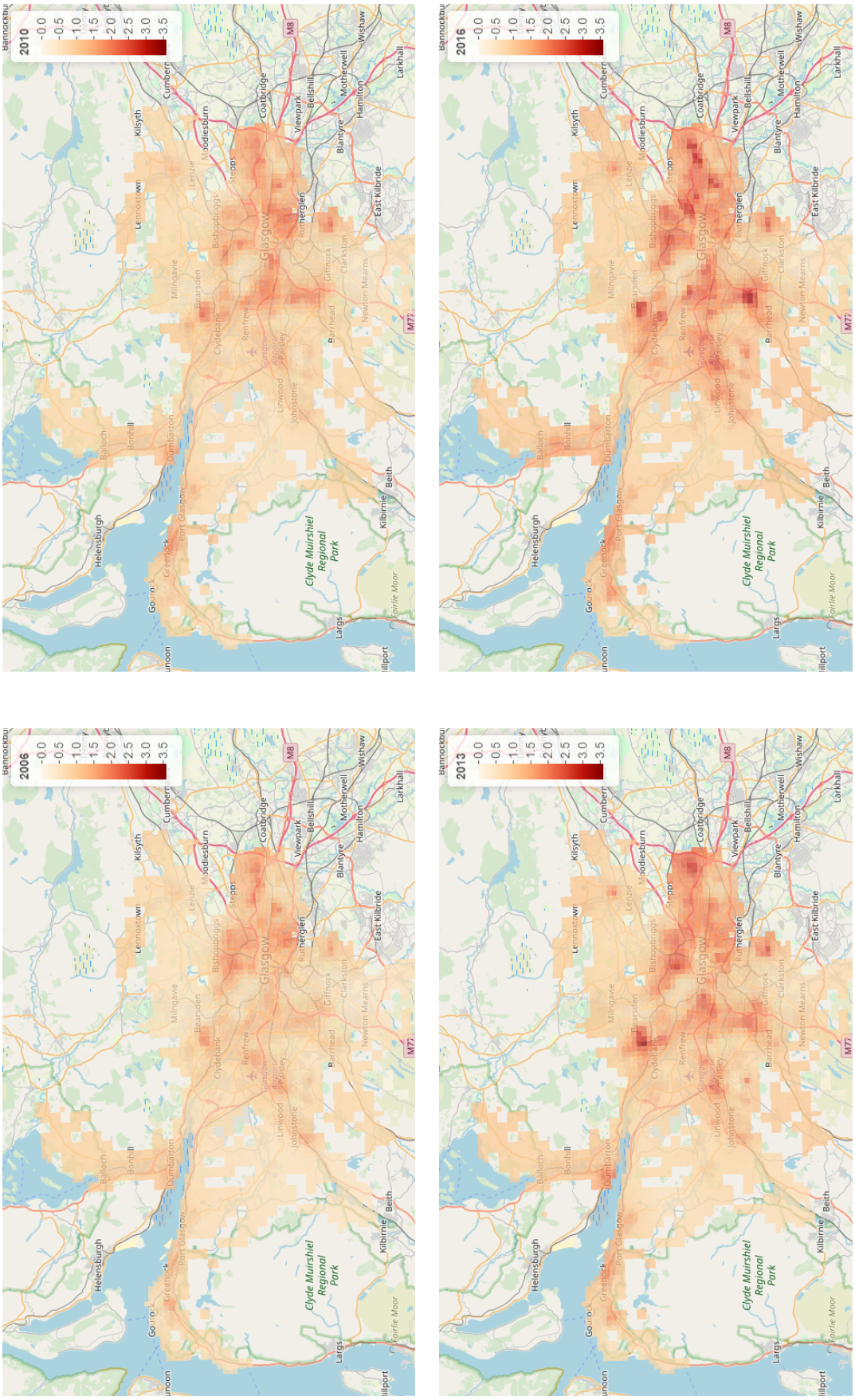

Figure 5.10: Spatial risk maps for respiratory disease for the years 2006, 2010, 2013, 2016.

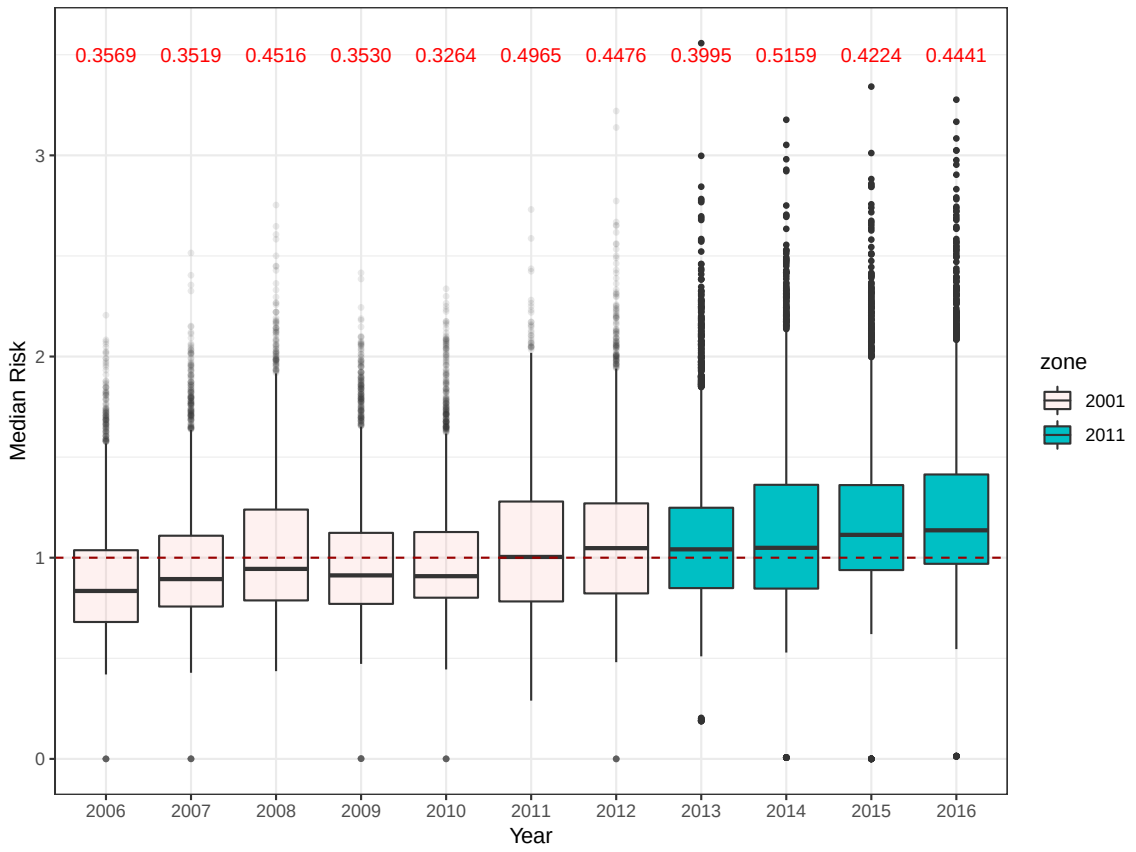

Figure 5.11: Boxplots of disease risk for respiratory disease in grids in Greater Glasgow and Clyde from 2006 - 2016. The IQR across grids are printed in red. Outliers are those observations that lie outside 1.5(IQR).

over time but the number of outliers with high risk also seems to have increased over time. In fact, in 2006 the grid with the highest risk of respiratory disease had an estimate of 2.205, meaning that those living in this grid are 2.2 time more at risk of respiratory disease than average. In 2016 this increases to 3.277, i.e. those living in this grid are more than 3 times more at risk of respiratory disease on average. These results are in line with what was found in Chapter 4 and show that the increase in health inequality which was found in Figure 4.9 until 2012 has continued to increase until 2014 where it may begin to level off slightly.

When comparing Figure 5.11 to Figure 5.5, which shows the boxplots of the raw SIR for all IGs over the time period, several differences may be noticed. Firstly, the median risk level in Figure 5.11 has been shifted downwards for every year. Initially this may seem odd as the median estimated risk of respiratory disease risk should not be significantly different from the median for the raw SIR, however we believe there is an explanation for this. Fundamentally these two figures are not showing the same thing, one is SIR based on the original IGs and the other shows estimated risk

on the grids. When looking at Figure 5.6 which shows the spatial pattern of SIR it can be seen that the areas with low values of SIR tend to be the geographically large IGs, whereas the areas with high values of SIR are predominantly the geographically small areas. Therefore, when the IG data is imputed onto the common grid, the areas which exhibit high risk, which are geographically small, have fewer grids to allocate the hospital admissions to. Compare this to the areas which exhibit low risk and are generally larger geographically, there are now a large number of grids to allocate a relatively small number of hospital admissions to, and so the proportion of grids which are estimated to have low risk will be much larger than the proportion of IGs which have low values of SIR. This will then lead to the median risk level being lowered when moving from IGs to grids. This phenomenon could also explain the other notable difference between Figures 5.5 and 5.11, which is the increased number of outliers with high estimated risk. To emphasise this point, the IG in Greater Glasgow and Clyde which is smallest geographically has an area of $0.19km^2$, which is in fact smaller than the area of each grid, which is $0.25km^2$. Compare this to the IG which is largest geographically, which has an area of $111.41km^2$. This is over 400 times larger than each grid square. Clearly then the number of grid squares which contain a considerable number of hospital admissions will be a substantially smaller proportion than at the IG level and so they are more likely to show up as outliers exhibiting high risk in Figure 5.11.

5.6 Discussion

In this chapter two multiple imputation approaches were proposed to estimate data on a common grid which allows for comparable inference over time when the boundaries on which data are collected on change. A simulation study was conducted to compare the two methods and the results showed than in terms of bias, RMSE and 95% coverage the posterior risk averaging approach (see Section 5.3.3) performed better than the data averaging approach (see Section 5.3.2). The posterior risk averaging approach was then applied to data containing hospital admissions for respiratory disease for the years 2006-2016 for the health board Greater Glasgow and Clyde,

where the data from 2013-2016 are reported on the redrawn IGs. A spatio-temporal model was then applied to the imputed data, $\hat{Y}_t(\mathcal{G}_i)$, to investigate how the risk of respiratory disease has changed over time across Greater Glasgow and Clyde and how health inequalities in risk have changed.

The main results from this chapter show that there has been an increase in risk of respiratory disease risk in Greater Glasgow and Clyde from 2006 - 2016. This is in line with findings from Chapter 4, however the disease risk has now been estimated for a more recent time period, during which the boundaries of the IGs which data are collected on were changed. This approach has therefore allowed for the modelling of this data which would not have been possible using the data in the original IG form. It was also found that the health inequality in respiratory disease risk has also increased over the time period, meaning that the differences in risk between Greater Glasgow and Clyde's most affluent and most deprived areas is getting worse. Again, this in the line with findings from Chapter 4 and shows that health inequalities have continued to increase from 2012 onwards.

Although the methods developed in this chapter allow for spatio-temporal modelling of data over a time period that would not have been possible using the original data, there is one clear drawback. Namely, if the Kriging of the SIR values onto the grid is not accurate, then the data on the grid which are used in the final modelling and hence the resulting inference will also not be accurate. Future work could therefore consider a data augmentation approach which would overcome this issue by combining the multinomial imputation step in Equation 5.5 within the McMC estimation algorithm which estimates the other parameters in the model. More specifically, the model parameters from 5.12, 5.13 and 5.14 could be estimated using McMC updates based on the current value of each $Y_t(\mathcal{G}_i)$. Then each $Y_t(\mathcal{G}_i)$ could be updated via data augmentation using the multiple imputation step 5.5 with the same weights as before (5.9). Here, $\hat{\theta}(\mathcal{G}_i)$, would be estimated within the McMC sampler rather than outwith the modelling as in our approach.

Chapter 6

Discussion and future work

This thesis focused on quantifying health inequalities across Scotland and estimating how they are changing over time, an issue which has huge repercussions for the people living in Scotland. There have been many studies on health inequalities in Scotland which were discussed in Chapter 1, however much of the existing research lacks in-depth analysis of health inequalities at the small area scale in Scotland, which is the focus of this thesis. Such approaches assess the extent and pattern of disease risk by partitioning the study region into a set of contiguous areal units, estimating the disease risk for each area and then presenting these risks on a disease map. The most common approaches are based on conditional autoregressive (CAR) models which are introduced in Chapter 2.

6.1 Single disease model

In Chapter 3, I proposed a spatio-temporal model for quantifying health inequalities in one disease in Scotland at a small area level using disease mapping techniques. This model was applied to data containing hospital admissions for one of Scotland's biggest killers, coronary heart disease (Scotpho, 2016), for the years 2003-2012. It was found that there are differences in coronary heart disease risk across Scotland and that these risks are changing over time. Overall, there was a decrease in coronary heart disease risk in Scotland over the time period. As well as seeing a decrease in risk over time, crucially, the health inequality in risk also decreased over the time period.

It was also of interest to identify any health inequalities in the risk of coronary heart disease between Scotland's 14 regional health boards and to estimate how they are changing over time. It was found that even after adjusting for deprivation (and other covariates), health inequalities in coronary heart disease risk still exist between the health boards although again, these have decreased over time.

6.2 Multi-disease model

In Chapter 4 the single disease model was extended to a more realistic multivariate disease risk model, where multiple diseases are investigated to better understand how health inequalities have changed across Scotland over time. This chapter proposed a novel spatio-temporal multi-disease model which was applied to data containing two more of Scotland's biggest killers, cerebrovascular disease and respiratory disease, alongside the coronary heart disease data. The between-disease correlation was estimated to be moderate between all pairs of disease which justifies our use of a multivariate modelling approach as it accounts for this correlation and allows for the diseases to borrow strength from each other. The results showed that although there was a decrease in risk for cerebrovascular disease and coronary heart disease overall in Scotland, this was accompanied by an increase in risk of respiratory disease. It was also found that across all IGs there was a decrease in health inequality for cerebrovascular disease and coronary heart disease over the time period, although they do still exist to a considerable extent. However, the inequalities in respiratory disease appear to be getting worse over the time period studied here. Similar to the results from Chapter 3, there was also still evidence of health inequalities between Scotland's health boards for all three diseases after covariate effects had been removed. Another concerning feature of these results are the similarities in the IGs which appear in the top 5 highest risk IGs at the start and end of the time period for each of the three diseases. Several IGs were placed in the top 5 in 2002 and 2012 for more than one disease. This shows the extent of the health inequality experienced in these areas and highlights that more needs to be done to target areas which are experiencing much higher risks of disease than the rest of Scotland.

6.3 Changing boundaries over time

One issue that was identified in this thesis and is a common problem in areal unit data of this type is a change to the boundaries during the time period for which data are collected. In 2014, the Scottish Government released a redrawn version of the intermediate geography boundaries. Therefore, the models which are outlined in Chapters 3 and 4 cannot be applied to any data which spans the time period in which this change occurs in the form that it is usually available, i.e. at IG level in Scotland. Chapter 5 therefore proposed two multiple imputation approaches, data averaging and posterior risk averaging (Sections 5.3.2 and 5.3.3), which address this problem by undertaking inference on a common grid for both sets of IGs, thus producing comparable inference over time. A simulation study was conducted to compare the two methods and the results showed that in terms of bias, RMSE and 95% coverage the posterior risk averaging approach permormed better than the data averaging approach. The posterior risk averaging approach was then applied to data containing hospital admissions for respiratory disease for the years 2006-2016 for the health board Greater Glasgow and Clyde, where the data from 2013-2016 are reported on the redrawn IGs. A spatio-temporal model was then applied to the imputed data to estimate how the risk of respiratory disease has changed over time in the Greater Glasgow and Clyde health board area and investigate any changes in health inequalities over this time period as well. The results from this model showed that the risk of respiratory disease has increased over the time period in Greater Glasgow and Clyde. This was accompanied by an increase in health inequalities in respiratory disease risk, indicating that the risk is increasing at a higher rate in areas whose risk was already elevated to begin with.

6.4 Common results and discussion

In this thesis, some common themes have been identified from Chapters 3, 4 and 5. First of all, health inequalities are improving for some diseases but not for others, demonstrating the importance of looking at more than one disease to get an over-

all picture of health inequalities in Scotland. In particular, it was found that the risk, and inequalities in risk, of coronary heart disease and cerebrovascular disease are decreasing over time across Scotland. However, the risk of respiratory disease as well as the health inequality in that risk are increasing over time across Scotland. Research into the prevalence of respiratory diseases in the UK by the British Lung Foundation has shown that in 2011 approximately 67% of UK hospital admissions from respiratory diseases were due to pneumonia, chronic obstructive pulmonary disease and acute lower respiratory infections. From these, both pneumonia, and more so chronic obstructive pulmonary disease, have shown increasing numbers of diagnoses over the time period 2004-2012 (The British Lung Foundation, 2013). It has also been estimated that the number of people with a chronic obstructive pulmonary disease diagnosis in Scotland will rise by 20% from 2011 to 2030 with a corresponding increase in costs of £48million (McLean et al., 2016). This increase in risk of respiratory disease, therefore, may be partially due to increased diagnoses, which in turn would result in an increase in hospital admissions. The increase in respiratory disease admissions could also be due to a reduction in competing causes for hospitalisation, given that there has been a reduction in risk of hospitalisation for cerebrovascular disease and coronary heart disease. Another important result is that this increase in hospital admissions is not occurring uniformly across all IGs. Instead, IGs which already had high risks of respiratory disease at the start of the time period are seeing more of an increase in numbers, which is driving the increase in the health inequality for this disease.

In Chapters 3 and 4 it was also found that these health inequalities do not just exist at the small area level but also between Scotland's 14 national health boards. The extent to which these exist is concerning as ideally the health board in which a person lives would not have an impact on their risk of disease, particularly at a large area level such as this. However, it was found that these health inequalities between the health boards have reduced quite significantly for cerebrovascular disease and coronary heart disease.

There were many significant changes to the structure of the health service in Scotland in the years before and during the time period of this data, which could help

to explain some of the improvements observed. In 1997, Designed to care: renewing the NHS in Scotland (Scottish Office, 1997) was published with the main aim of phasing out the internal market, integrating services to eliminate duplication and wasteful competition and merging 47 Scottish trusts into 28. Following the Scottish devolution in 1999, Our national health: a plan for action, a plan for change (Scottish Executive, 2000) stated that the health budget in Scotland was due to rise from £4.9 billion in 1999-2000 to £6.7 billion in 2003-2004 (which coincides with the start of the time-period in Chapters 3 and 4). This considerable resource increase was to be used to build a modernised health system and improve the health of Scotland's population.

In the period 2003-2006 there were several more policy changes which may have contributed to the results found in this thesis. First of all, in 2003, Partnership for care: Scotland's health White Paper (NHS Scotland, 2003b) was released, which led to the abolition of the NHS Trusts in 2004. These were absorbed into the health boards which have been the focus of this thesis. The health boards were given single tier responsibility for governance and accountability and health improvement was made a priority. This policy change also led to the creation of 40 community health partnerships (CHPs) which were the vehicle for planning and delivery of primary and community based services (The Scottish Government, 2010). The CHPs were given the responsibility (along with the health boards) in improving health and reducing health inequalities and were able to work locally, not only to tackle smoking, obesity, drug and alcohol misuse etc, but also to work with other agencies to tackle many of the other social determinants of health.

The National Service Framework (Scottish Executive, 2005) published in 2005 set out long-term plans for the NHS in Scotland over the next 20 years. The key message was to look to the population of Scotland to take more responsibility for their own health and '*anticipate and prevent rather than react*'. Another key piece of legislation during this time period was The Smoking, Health and Social Care (Scotland) Act 2005, which banned smoking in any enclosed public space in Scotland from 26 March 2006. The ban was described by the Chief Medical Officer, Mac Armstrong as bringing '*far and away the most important improvement in our health in a generation*'. In

2007, Better Health, Better Care: Action Plan (The Scottish Government, 2007) was introduced which views patients and the public as '*partners rather than recipients of care*'. This again outlines the focus of helping the public improve their health, acting proactively rather than reactively, particularly in disadvantaged communities. This shift in perspective, I believe, is crucial to truly tackling the health inequalities that still exist in Scotland.

Reducing the size of Scotland's health inequalities has clearly been a key focus of both the Scottish Government and NHS Scotland, with many of these policy changes directly stating that an improvement in health inequalities as well as overall health is of importance for Scotland's people moving forward. This thesis quantifies these changes, and a concerning feature of the results is the large number of outliers with high risk estimates as illustrated in Figures 4.9 and 5.11. This further highlights the huge problem that Scotland faces in inequality in the overall health of its population and that more needs to be done to target areas which are experiencing much higher risks of disease than other parts of Scotland.

6.5 Limitations and future work

There are several limitations to this work, some of which were unavoidable while others only came to light as a result of the research undertaken. From a data perspective, ideally there would have been data available for a longer time period to allow for more complex modelling of the temporal aspect to the models in Chapters 3 and 4. However, as described in Chapter 5, any data collected after 2012 could not have been used in its original form. In Chapter 5 if possible, this approach would have been implemented over the same study region as the previous two chapters, i.e. over all of Scotland. However considering there were 3137 grid squares in Greater Glasgow and Clyde alone, over 11 years this leads to 34507 data points. To do this over all of Scotland at this scale would have been computationally infeasible and so it was decided to apply the model to one health board only. This also means we were unable to compare health inequalities between the health boards as has been done in Chapters 3 and 4. Future work could consider applying this method to all of Scotland

using a lattice of cells of a larger area. However, this would be a trade-off between having grid cells which are large enough to cover all of Scotland without resulting in an infeasibly large data set whilst ensuring that the grids are small enough to provide the information needed to investigate health inequalities at the small area level, and that the resulting data aren't aggregated too much. For example, in areas where the original IGs are small (e.g. in cities), a lattice of larger grid cells would lead to loss of information as disease counts in neighbouring IGs would be merged together. A way to overcome this could utilise an adaptive lattice with non-uniform grid cell size.

The methods developed in Chapter 5 has one main drawback in that if the imputation of $\hat{\theta}(\mathcal{G}_i)$ is not accurate, then the data, $Y_t(\mathcal{G}_i)$,which the spatio-temporal model is fitted to and hence the resulting inference will be inaccurate. Therefore, future work could develop an approach which overcomes these issues, for example, a data augmentation strategy which allows for $Y_t(\mathcal{G}_i)$ to estimated within the McMC estimation algorithm.

It may also be of value to implement these methods using a multivariate approach similar to that proposed in Chapter 4 to allow for a better understanding of how health inequalities have changed across Greater Glasgow and Clyde. However, as yet, there are no freely available data over this time period for coronary heart disease and cerebrovascular disease. Developing a data augmentation technique within the multivariate model proposed in Chapter 4 and applying this to data for multiple diseases when these datasets are freely available would be a worthwhile future project.

Appendix A

Statistical properties

A.1 Conditional Distribution Property of a Multivariate Gaussian Distribution

Given the joint distribution of $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right). \tag{A.1}$$

Then the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is given by

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \tag{A.2}$$

Appendix B

Comparison of the multivariate model from Chapter 4 to other models

B.1 Temporally changing beta

The results from the model which allows the regression parameters to change over time can be found seperately for each disease in Tables B.1, B.2 and B.3. In general these show little change in both the point estimates and the 95% credible intervals over time and compared with the corresponding estimates from the model with temporally static estimates (Table 4.2). In almost all cases the 95% credible intervals overlap for all pairs of time periods.

B.2 No covariates

In order to test the sensitivity of our results to the choice of covariates, our model was run again with no covariate effects. The results were virtually the same and some figures are shown here for comparison. Figure B.1 shows boxplots of the disease risk for all three diseases from the model with no covariates. When compared to Figure 4.9 in the main text, it can be seen that there is very little change. Figure B.2 in

Table B.1: Relative risk estimates for a 1% increase in each covariate (not urban/rural covariate) and 95% credible intervals for the covariates in a model with temporally varying regression parameters for cerebrovascular disease. Significant results are in bold.

Covariate - Cerebrovascular	**Median RR**	**95% CI**
% 16-64 year olds claiming JSA		
2003	**1.063**	**(1.054, 1.072)**
2004	**1.061**	**(1.052, 1.069)**
2005	**1.065**	**(1.057, 1.074)**
2006	**1.059**	**(1.050, 1.067)**
2007	**1.058**	**(1.049, 1.067)**
2008	**1.061**	**(1.053, 1.070)**
2009	**1.059**	**(1.051, 1.068)**
2010	**1.056**	**(1.048, 1.065)**
2011	**1.056**	**(1.047, 1.065)**
2012	**1.060**	**(1.051, 1.068)**
Log % Asian		
2003	1.004	(0.980, 1.029)
2004	0.990	(0.967, 1.014)
2005	1.004	(0.979, 1.027)
2006	1.015	(0.992, 1.040)
2007	0.985	(0.962, 1.009)
2008	0.993	(0.970, 1.016)
2009	1.000	(0.981, 1.029)
2010	**0.995**	**(0.940, 0.987)**
2011	0.993	(0.968, 1.017)
2012	1.018	(0.994, 1.043)
Log % Black		
2003	1.005	(0.992, 1.017)
2004	**1.023**	**(1.010, 1.036)**
2005	**1.014**	**(1.002, 1.027)**
2006	1.002	(0.989, 1.014)
2007	**1.017**	**(1.002, 1.031)**
2008	1.001	(0.989, 1.013)
2009	1.008	(0.995, 1.021)
2010	**1.018**	**(1.006, 1.031)**
2011	1.009	(0.996, 1.022)
2012	0.992	(0.980, 1.004)
Rural area		
2003	1.036	(0.978, 1.098)
2004	1.009	(0.954, 1.068)
2005	0.997	(0.940, 1.053)
2006	**0.938**	**(0.888, 0.992)**
2007	0.958	(0.903, 1.011)
2008	0.960	(0.910, 1.013)
2009	0.993	(0.940, 1.046)
2010	0.950	(0.902, 1.006)
2011	1.004	(0.950, 1.060)
2012	**0.940**	**(0.887, 0.994)**

Table B.2: Relative risk estimates for a 1% increase in each covariate (not urban/rural covariate) and 95% credible intervals for the covariates in a model with temporally varying regression parameters for coronary heart disease. Significant results are in bold.

Covariate - Coronary Heart Disease	Median RR	95% CI
% 16-64 year olds claiming JSA		
2003	**1.067**	**(1.059, 1.074)**
2004	**1.067**	**(1.060, 1.075)**
2005	**1.063**	**(1.055, 1.070)**
2006	**1.058**	**(1.050, 1.065)**
2007	**1.059**	**(1.052, 1.067)**
2008	**1.069**	**(1.061, 1.076)**
2009	**1.066**	**(1.059, 1.074)**
2010	**1.065**	**(1.057, 1.073)**
2011	**1.063**	**(1.055, 1.071)**
2012	**1.067**	**(1.061, 1.077)**
Log % Asian		
2003	**0.944**	**(0.927, 0.962)**
2004	**0.953**	**(0.936, 0.971)**
2005	**0.932**	**(0.916, 0.950)**
2006	**0.944**	**(0.927, 0.962)**
2007	**0.961**	**(0.944, 0.980)**
2008	**0.977**	**(0.960, 0.998)**
2009	1.002	(0.983, 1.010)
2010	0.989	(0.970, 1.009)
2011	**0.974**	**(0.956, 0.995)**
2012	0.985	(0.966, 1.007)
Log % Black		
2003	1.003	(0.994, 1.013)
2004	1.003	(0.994, 1.013)
2005	0.999	(0.989, 1.008)
2006	0.996	(0.986, 1.005)
2007	1.000	(0.991, 1.010)
2008	0.997	(0.988, 1.007)
2009	1.000	(0.990, 1.009)
2010	1.006	(0.996, 1.017)
2011	1.004	(0.994, 1.015)
2012	1.012	(1.002, 1.022)
Rural area		
2003	**0.920**	**(0.881, 0.965)**
2004	**0.942**	**(0.901, 0.984)**
2005	**0.911**	**(0.871, 0.955)**
2006	**0.942**	**(0.900, 0.984)**
2007	**0.954**	**(0.913, 0.997)**
2008	0.968	(0.925, 1.012)
2009	0.978	(0.935, 1.024)
2010	0.989	(0.945, 1.034)
2011	0.980	(0.934, 1.028)
2012	0.987	(0.942, 1.033)

Table B.3: Relative risk estimates for a 1% increase in each covariate (not urban/rural covariate) and 95% credible intervals for the covariates in a model with temporally varying regression parameters for respiratory disease. Significant results are in bold

Covariate - Respiratory	**Median RR**	**95% CI**
% 16-64 year olds claiming JSA		
2003	**1.100**	**(1.093, 1.078)**
2004	**1.101**	**(1.094, 1.110)**
2005	**1.104**	**(1.097, 1.113)**
2006	**1.101**	**(1.095, 1.109)**
2007	**1.101**	**(1.094, 1.109)**
2008	**1.108**	**(1.100, 1.116)**
2009	**1.108**	**(1.101, 1.116)**
2010	**1.104**	**(1.098, 1.113)**
2011	**1.104**	**(1.097, 1.112)**
2012	**1.106**	**(1.100, 1.114)**
Log % Asian		
2003	**0.970**	**(0.953, 0.988)**
2004	0.986	(0.969, 1.005)
2005	0.987	(0.971, 1.006)
2006	0.983	(0.966, 1.001)
2007	**0.977**	**(0.960, 0.995)**
2008	**0.979**	**(0.961, 0.996)**
2009	1.000	(0.982, 1.017)
2010	0.995	(0.978, 1.013)
2011	**0.976**	**(0.960, 0.994)**
2012	**0.981**	**(0.963, 0.999)**
Log % Black		
2003	0.999	(0.990, 1.008)
2004	0.997	(0.988, 1.005)
2005	0.993	(0.985, 1.001)
2006	0.998	(0.990, 1.006)
2007	0.997	(0.989, 1.004)
2008	0.992	(0.984, 1.000)
2009	**0.987**	**(0.979, 0.995)**
2010	**0.991**	**(0.983, 0.999)**
2011	0.992	(0.984, 1.000)
2012	**0.988**	**(0.980, 0.996)**
Rural area		
2003	0.993	(0.957, 1.040)
2004	1.011	(0.970, 1.055)
2005	1.015	(0.977, 1.058)
2006	1.027	(0.989, 1.072)
2007	0.991	(0.954, 1.030)
2008	0.968	(0.971, 1.051)
2009	0.978	(0.987, 1.069)
2010	0.989	(0.936, 1.010)
2011	0.980	(0.927, 1.000)
2012	0.987	(0.964, 1.041)

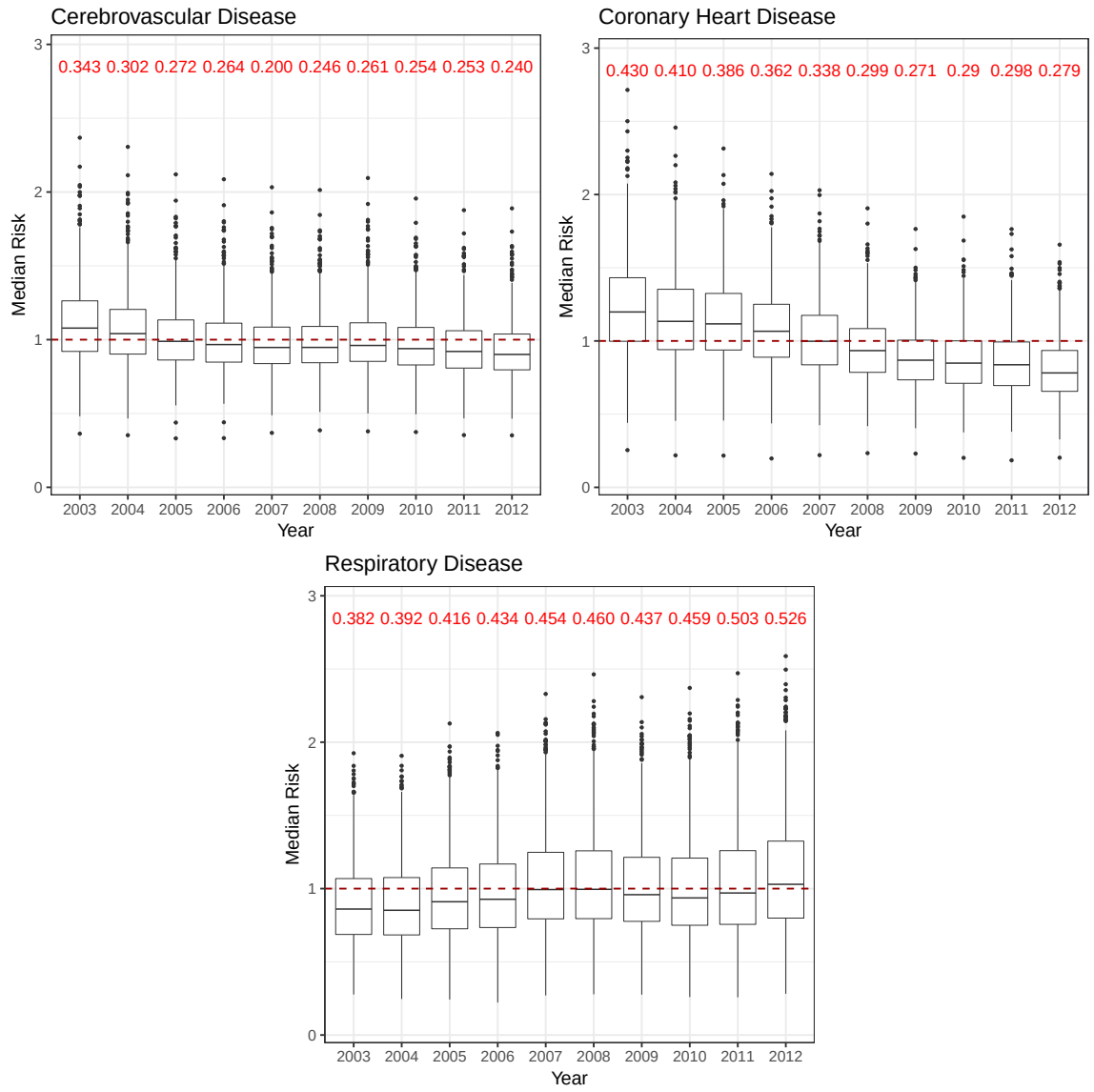

Figure B.1: Boxplots of disease risk for a model without covariates for cerebrovascular disease, coronary heart disease, and respiratory disease in IG's in Scotland from 2003 - 2012. The IQR across IG's are printed in red. Outliers are those observations that lie outside 1.5*IQR.

this supplementary material shows the fitted values from the model with covariates (x-axis) and the model without (y-axis) and the results are practically unchanged.

B.3 Multivariate spatio-temporal random effect

Figure B.3 shows the disease risk for each of the diseases in IG's in Scotland across the time period using the results from the model proposed by Quick et al. (2017b). When compared to the same plot using the results from model 4.1 (Figure 4.9) it can be seen that these results are practically identical. This is backed up in Figure B.4, which shows the fitted values from our model (4.1) on the x-axis and the corresponding fitted values from the model proposed by Quick et al. (2017b) (4.6) on the y-axis.

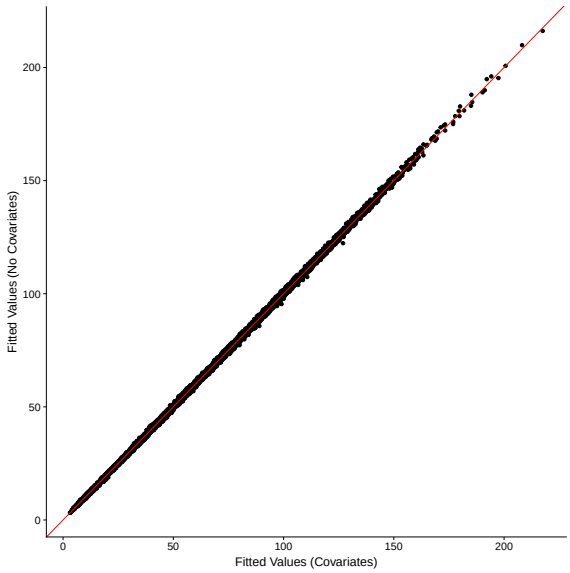

Figure B.2: Scatterplot of fitted values from model with covariates vs fitted values from model without covariates.

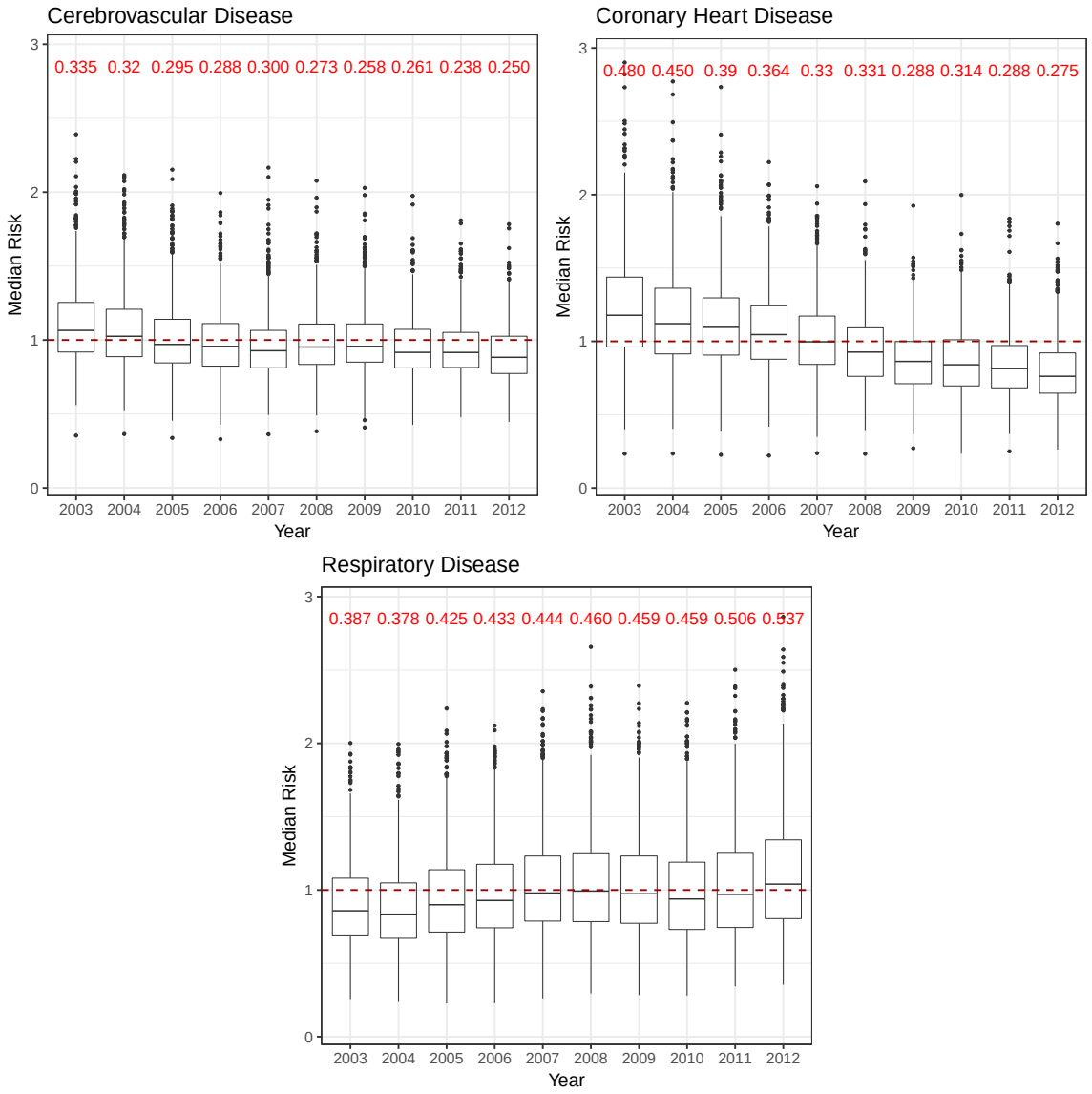

Figure B.3: Boxplots of disease risk from the Quick et al. (2017b) model for cerebrovascular disease, coronary heart disease, and respiratory disease in IG's in Scotland from 2003 - 2012. The IQR across IG's are printed in red. Outliers are those observations that lie outside 1.5*IQR.

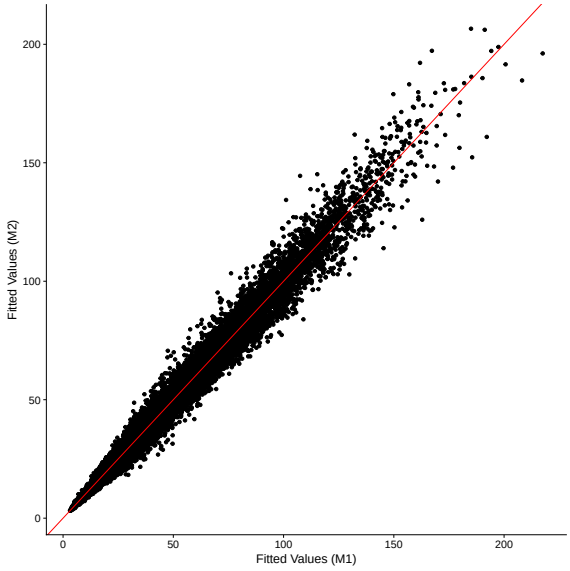

Figure B.4: Scatterplot of fitted values from our model (4.1) vs fitted values from the Quick et al. (2017b) model 4.6.

Again this shows that the results are very similar.

References

Acheson, D. (1998). *Independent Inquiry into Inequalities in Health: Report.* HMSO, London. ISBN 9780113221738. 2

Audit Scotland (2012). Health Inequalities in Scotland. http://www.audit-scotland.gov.uk/docs/health/2012/nr_121213_health_inequalities.pdf. 2016-11-16. 3, 6, 45, 48

Bartley, M. (2016). *Health inequality: an introduction to theories, concepts and methods.* Polity, Cambridge, 2nd edition. ISBN 978-0-745-69109-1. 2

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370 – 418. 14

Beeston, C., McCartney, G., Ford, J., Wimbush, E., Beck, S., MacDonald, W., and Fraser, A. (2013). Health Inequalities Policy review for the Scottish Ministerial Task Force on Health Inequalities. NHS Scotland. Edinburgh. 3

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443. 29, 31

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1 – 20. 26, 51, 79

Black, D., Morris, J., and Townsend, P. (1982). *Inequalities in Health: The Black Report and the Health Divide.* Penguin, Harmondsworth. ISBN 978-0140172652. 1, 2

Centers for Disease Control and Prevention (2014). Smoking and respiratory diseases. https://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/pdfs/fs_smoking_respiratory_508.pdf. 2017-08-16. 92

Diggle, P., Moraga, P., Rowlingson, B., and Taylor, B. (2013). Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. *Statistical Science*, 28(4):542–563. 38

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics*. Springer Series in Statistics, 1st ed edition. 22, 23

Dobson, A. J. and Barnett, A. G. (2008). *An introduction to generalized linear models*. CRC Press, London, 3rd edition. 22

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7. 52, 80

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18. 52, 80

Flenberg, S. E. (2006). When did bayesian inference become 'bayesian'? *Bayesian Anal*, 1:1 – 40. 14

Fotheringham, A. S. and Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23:1025–1044. 37

Garthwaite, P., Joliffe, I., and Jones, B. (2006). *Statistical Inference*. OUP, Clarendon Street, Oxford, 2nd edition. 13

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4:11–25. 34, 36

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

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721 – 741. 16

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.* University Press. 19, 55, 82, 119

Glasgow Open Data (2010). Life expectancy figures in males. 2016-05-15. 3

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97 – 109. 16

Hemmingway, H. and Marmot, M. (1999). Psychosocial factors in the aetiology and prognosis of coronary heart disease: systematic review of prospective cohort studies. *Br. Med. J.*, 318:1460–7. 2

Heywood, I. D., Cornelius, S., and Carver, S. (1998). *An introduction to geographical information systems.* Addison Wesley Longman, New York. 37

Jelinski, D. E. and Wu, J. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11(3):129–140. 37

Kim, H., Sun, D., , and Tsutakawa, R. K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96:1506–1521. 33

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567. 30, 32

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, 17(18):2045–2060. 31

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6):119–139. 23

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1 – 14. 22

Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-temporal areal unit modelling in R with conditional autoregressive priors using the CARBayesST package. *Journal of Statistical Software*, 84(9). 114

Leroux, B., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for for spatial dependence. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 135–78. Springer-Verlag, New York. 28, 50, 79, 113

Leyland, A. H., Dundas, R., McLoone, P., and Boddy, F. A. (2007). Cause-specific inequalities in mortality in Scotland: two decades of change. A population-based study. *BMC Public Health*, 7(1):172. 5

Li, Y., Brown, P., Gesink, D., and Rue, H. (2012a). Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research.*, 21(5):479–507. 37, 38, 100

Li, Y., Brown, P., Rue, H., al Maini, M., and P.Fortin (2012b). 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.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3:112 – 115. 19

MacNab, Y. and Dean, C. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21(3):347–358. 29

MacNab, Y. C. (2016). Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate CAR models. *Statistics in Medicine*, 35:3827–3850. 34

Marmot, M. (2005). Social determinants of health inequalities. *Lancet*, 365:1099–104.

Marmot, M. (2010). *Fair society, healthy lives : the Marmot Review : strategic review of health inequalities in England post-2010.* Institute of Health Inequity, University College London. ISBN 9780956487001. 2

McCartney, G. (2012). What would be sufficient to reduce health inequalities in Scotland? *Ministerial Task Force on Health Inequalities*, (12):Paper 3(a). 1, 11

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,*. Chapman and Hall/CRC, London, 2nd edition. 21

McLean, S., Hoogendoorn, M., Hoogenveen, R., Feenstra, T., Wild, S., Simpson, C., van Mölken, M. R., and Sheikh, A. (2016). Projecting the COPD population and costs in england and scotland: 2011 to 2030. *Scientific Reports*, 6. 128

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23. 25, 49, 74

Napier, G., Lee, D., Robertson, C., Lawson, A., and Pollock, K. (2016). A Model to Estimate the Impact of Changes in MMR Vaccination Uptake on Inequalities in Measles Susceptibility in Scotland. *Statistical Methods in Medical Research*, 25:1185–1200. 112, 117

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, 135(3):370 – 384. 20

NHS Health Scotland (2015). 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. 2017-10-10. 3

NHS Scotland (2003a). Health in Scotland. http://www.gov.scot/Resource/Doc/47237/0013499.pdf. 2017-08-16. 92

NHS Scotland (2003b). Partnership for care: Scotland's health White Paper. http://www.gov.scot/Resource/Doc/47032/0013897.pdf. 2018-07-12. 129

Popham, F. and Boyle, P. (2011). Assessing socio-economic inequalities in mortality and other health outcomes at the Scottish national level. http://www.scphrp.ac.uk/wp-content/uploads/2014/05/Assessing-socio-economic-inequalities-in-mortality-and-other-health-outcomes.pdf. 2017-10-10. 3

Quick, H., Waller, L. A., and Casper, M. (2017a). A multivariate space time model for analysing county level heart disease death rates by race and sex. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 36

Quick, H., Waller, L. A., and Casper, M. (2017b). Multivariate spatiotemporal modeling of age-specific stroke mortality. *The Annals of Applied Statistics*, 11(4):2170 – 2182. xiv, 37, 94, 96, 137, 138, 139

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 52, 80, 114

Reis, S., Steinle, S., Carnell, E., Leaver, D., Vieno, M., Beck, R., and Dragosits, U. (2015). UK gridded population based on Census 2011 and land cover map 2007. 102

Ribeiro Jr, P. and Diggle, P. (2001). geor: A package for geostatistical analysis. *R-NEWS*, 1(2). 110

Richardson, S., Abellan, J. J., and Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical Methods in Medical Research*, 15:385 – 407. 35

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351 – 367. 18

Rostami, M., Mohammadi, Y., and Jalilian, A. (2017). Modeling spatio-temporal variations of substance abuse mortality in Iran using a log-Gaussian Cox point process. *Spatial and Spatio-temporal Epidemiology*, 22:15–25. 39

Rushworth, A., Lee, D., and Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, 10:29–38. 32

Scotpho (2016). Deaths: most frequent causes. http://www.scotpho.org.uk/population-dynamics/deaths/data/most-frequent-causes. 2016-04-12. 6, 40, 67, 125

Scottish Executive (2000). Our National Health: a plan for action, a plan for change. http://www.gov.scot/Resource/Doc/158732/0043081.pdf. 2018-07-12. 129

Scottish Executive (2005). A National Framework for Service Change in the NHS in Scotland. http://www.sehd.scot.nhs.uk/nationalframework/Reports.htm. 2018-07-12. 129

Scottish Office (1997). Designed to Care: Renewing the National Health Service in Scotland. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/260828/scotnhs.pdf. 2018-07-12. 129

Stern, H. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19(17-18):2377 – 2397. 27

Stroke Association (2016). State of the Nation. https://www.stroke.org.uk/sites/default/files/stroke_statistics_2015.pdf. 2017-08-16. 93

Taulbut, M., Walsh, D., McCartney, G., Parcell, S., Hartmann, A., Poirier, G., Strniskova, D., and Hanlon, P. (2014). Spatial inequalities in life expectancy within postindustrial regions of Europe: a cross-sectional observational study. *British Medical Journal Open*, 4(6). 5

Taylor, B., Andrade-Pacheco, R., and Sturrock, H. (2017). Continuous Inference for Aggregated Point Process Data. *Journal of the Royal Statistical Society. Series A*, 181(4):1125–1150. 38

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatio-temporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software*, 63(7). 38

The British Lung Foundation (2013). Lung disease in the UK - big picture statistics. https://statistics.blf.org.uk/lung-disease-uk-big-picture. 2017-08-17. 128

The Scottish Government (2007). Better Health, Better Care: Action Plan. http://www.gov.scot/Resource/Doc/206458/0054871.pdf. 2018-07-12. 130

The Scottish Government (2008). Equally Well: report of the ministerial task force on health inequalities. http://www.gov.scot/Resource/Doc/229649/0062206.pdf. 2017-10-10. 3

The Scottish Government (2010). Study of Community Health Partnerships. https://www2.gov.scot/Publications/2010/05/06171600/0. 2018-07-12. 129

The Scottish Government (2012). The Scottish Health Survey: Equality Groups. http://www.gov.scot/Resource/0040/00406749.pdf. 2016-10-12. 45, 64

The Scottish Government (2015). Review of Equality Evidence in Rural Scotland. http://www.gov.scot/Resource/0046/00469898.pdf. 2016-10-12. 64

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–40. 119

Tzala, E. and Best, N. (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, 17:97 – 118. 35

Ugarte, M., Etxeberria, J., Goicoa, T., and Ardanaz, E. (2012). Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990-2005). *Cancer Epidemiology*, 36(3):270–289. 32

U.S Department of Health and Human Services (2014). The Health Consequences of Smoking - 50 Years of Progress. 41

Walsh, D., McCartney, G., Collins, C., Taulbut, M., and Batty, G. D. (2016). History, politics and vulnerability: explaining excess mortality in Scotland and Glasgow. http://www.gcph.co.uk/assets/0000/5586/History_politics_and_vulnerability.pdf. 2017-10-10. xi, 2, 4, 5

World Health Organization (2008). Social determinants of health. http://www.who.int/social_determinants/thecommission/finalreport/key_concepts/en/. 2017-09-26. 1

World Health Organization (2016). Life expectancy increased by 5 years since 2000, but health inequalities persist. http://www.who.int/mediacentre/news/releases/2016/health-inequalities-persist/en/. 2017-10-31. 1