Spatial modelling of air pollution, deprivation and mortality in Scotland

Francesca Giuseppina Pannullo

Submitted in fulfilment of the requirement for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow
September 2017

'Plans to protect air and water, wilderness and wildlife are in fact plans to protect man.'

— Stewart Udall

Abstract

Air pollution is not only a major risk to the environment, but also a major environmental risk to the health of the population in developed and developing countries. The health impact of both short-term and long-term exposure to air pollution has been the focus of much research in the past few decades, which has investigated the relationship between specific air pollutants, such as carbon monoxide (CO), nitrogen dioxide (NO_2), particulate matter ($PM_{2.5}$ and PM_{10}), and sulphur dioxide (SO_2), to cardiovascular and respiratory diseases.

The health impact of short-term exposure is conducted through time series studies, whereas long-term exposure is investigated through cohort studies. Cohort studies are considered the gold-standard research design since inference is made at the individual level and can directly assess cause and effect. However, cohort studies are costly and require a long follow-up period meaning they take a long time to conduct.

To counteract these limitations, spatial ecological studies are used instead, which make use of routinely available disease data and air pollutant concentrations at a small areal level, such as census tracts or postcodes. This is to ensure the population under study is relatively homogeneous within the areal unit in terms of socio-demographic characteristics, and thus complements inference from a cohort study. These studies quantify the health impact of exposure to air pollution by relating geographical contrasts between air pollutant concentrations and disease risk across the chosen spatial resolution. The disease data are counts of the numbers of disease cases occurring in each areal unit, and Poisson log-linear models are used to assess the pollutant-health relationship.

Other covariate information, such as socio-economic deprivation, is also included to help explain the spatial pattern in disease risk. However, the residual disease risk after the covariate effects have been accounted for tends to contain spatial autocorrelation, which has to be modelled in order to make sound inferences. Residual spatial autocorrelation is typically modelled by a set of random effects that utilise a neighbourhood matrix in order to induce spatial autocorrelation into the model. There are a number of specifications to model this, but this thesis makes use of the Leroux specification due to its flexibility in being able to model both strong and weak spatial autocorrelation.

An important issue with using a spatial ecological study design is the estimation of spatially representative pollutant concentrations that are available in each areal unit. Studies can typically use measured data from fixed-location monitors that are spatially sparse and do not provide a pollutant concentration for each areal unit; or they make use of modelled concentrations available at a fine grid square resolution, which are known to contain biases and no measure of uncertainty. There have been numerous statistical approaches to combine both sets of information in order to estimate accurate and spatially representative concentrations. This thesis will develop previous methodology that utilises extra data sources in order to improve the prediction performance of the model for use in a Scottish context.

The overarching aim of this thesis is to investigate the cardio-respiratory health effects of long-term exposure to air pollution in West Central Scotland, UK. As the majority of air pollution in this region results from vehicle emissions, nitrogen dioxide (NO_2), a traffic-related gaseous pollutant, will be used to measure air pollution. Models investigating its health effect will incorporate predicted measures of NO_2 developed in this thesis. The sensitivity of the pollutant-health effect to the choice of NO_2 concentrations, indicator of deprivation, and choice of spatial model will be investigated. Changing these factors has been shown to modify estimated pollutant-health effects.

Findings in this thesis demonstrated that improvements in the accuracy of fine scale spatial prediction of NO_2 concentrations can be made by utilising extra sources of data in addition to the commonly-used monitoring stations. In addition, the estimated pollutant-health effect is not robust to the choice of the aforementioned factors and the choice of these factors can have a major impact on the resulting pollutant-health effects. This justified the combination of all statistical models into a single effect size, which estimated a small, but positive effect of NO_2 concentrations on cardio-respiratory ill health. However, the estimated NO_2-health relationship was not substantial, possibly due to the NO_2 concentrations in West Central Scotland being too low. Greater variation in the exposure would be needed to observe substantial health impacts.

Acknowledgements

Firstly, I would like to acknowledge the Medical Research Council Social and Public Health Sciences Unit for funding this PhD, and the School of Mathematics and Statistics for their support and fantastic opportunities throughout these three years.

I would like to thank my supervisors: Professor Alastair H Leyland, Dr Duncan Lee, and Dr Eugene Waclawski for their help, guidance and support over the course of my PhD as it is through their belief in me that I have got to where I am today. I would also like to thank them for their friendship and all the other amazing opportunities that have come from being supervised by them. It has been a privilege working with such highly-respected academics.

I owe a big thank you to my mum, dad and brother who have constantly supported and believed in me especially when I didn't believe in myself. Without them this research would never have been completed. I especially want to thank my mum, Esther, for her amazing understanding in all aspects of my life as without her I wouldn't be the person I am today. Oscar, I know you are a dog and cannot read this, but I thank you for always being up for a long walk when I needed it most, and for being there when I needed cuddles.

I would like to thank Patricia for being the best office mum anyone could ask for, and to all the friends I have made within both the SPHSU and the School of Mathematics and Statistics for their continued support, especially to Cat for being my soundboard and proof-reader, and to the statistics PhD students for being awesome and at least once winning the Monday night pub quiz.

Finally, I would like to thank my best friends Gavin, Lauren, and Michael for adopting me into their lives, providing me with the best times I have ever had and for their continued support throughout every aspect of my life.

Declaration

I have prepared this thesis myself and no part of it has been submitted previously as part of any application for a degree. I carried out all aspects of the research, except where otherwise stated.

The research presented in Chapter 4 has been published in the Atmospheric Environment journal with the title *Improving spatial nitrogen dioxide prediction using diffusion tubes: A case study in West Central Scotland* (2015, volume 118, p227-235), and is co-authored with Dr Duncan Lee, Dr Eugene Waclawski, and Professor Alastair H Leyland. In addition, Chapter 5 has been published in the Spatial and Spatio-temporal Epidemiology journal with the title *How robust are the estimated effects of air pollution on health? Accounting for model uncertainty using Bayesian model averaging* (2016, volume 18, p53-62), and is also co-authored with Dr Duncan Lee, Dr Eugene Waclawski, and Professor Alastair H Leyland. Furthermore, I presented these works at the 9th international GEOMED conference in Florence, Italy in 2015, and the 10th European Public Health (EPH) conference in Vienna, Austria in 2016 and was published in the conference proceedings (The European Journal of Public Health) with the title *Accounting for model uncertainty due to deprivation in the study of air pollution and health effects* (2016, volume 26, suppl 1).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Air pollution and health

Air pollution has been a major public health concern for over 700 years; however, it only came to global prominence in the last 80 years and is the largest single environmental health risk today. It is estimated to kill 1 in 8 people globally and by reducing air pollution levels, many countries can reduce the burden of disease from chronic and acute respiratory diseases, such as asthma, stroke, lung cancer, and heart disease. It has repeatedly been shown to have a detrimental impact on human health, with some of the earliest prominent examples being the Meuse Valley in Belgium in 1930 (Firket, 1936); Donora, Pennsylvania in October 1948 (Ciocco & Thompson, 1961); the London smog episode of December 1952 (Ministry of Public Health, 1954); and more recently in Shanghai, Eastern China in December 2013 (Huang et al., 2016). The London episode resulted in more than 3,000 excess deaths (as displayed in Figure 1.1) compared with previous years and brought the harmful effects of air pollution to the forefront. These air pollution episodes were caused by industrial pollution sources and stagnant weather conditions, which caused a sharp increase in air pollutant concentrations over several days (Brunekreef & Holgate, 2002). For the London episode, a government committee was set-up to examine the effects of air pollution and what caused these episodes, and a report was produced that expressed an *'emphatic belief that air pollution on the scale with which we are familiar is a social and economic evil which should no longer be tolerated... To do this will require a national effort, will entail costs and sacrifices, and the recommendations made will involve expenditure by government, local authorities, industry and householders alike'* (Stewart, 1994).

High pollution episodes such as these have lead to the implementation of air pollution legislation, such as the Clean Air Acts in 1956 and 1993 in the UK; the UK Air Quality Strategy in 1997, 2000 and 2007 (Department for the Environment, Food and Rural Affairs, 2007); and the 2008 Ambient Air Quality Directive in the European Union (2008/50/EC). These policies have lead to a reduction in air pollution concentrations in many parts of the world. However, a recent report by the World Health

Organisation (WHO) estimated that outdoor air pollution contributed to 3.7 million premature deaths in people under the age of 60 in 2012 (World Health Organisation, 2014), and in 2014, 95% of the world's population was living in places where levels of air pollution were exceeding the WHO air quality guideline levels.

Air pollution remains a serious public health problem in the UK. Previous studies have reported associations between numerous air pollutants, such as carbon monoxide (CO, Villeneuve et al., 2003), nitrogen dioxide (NO_2, Bennett et al., 2014), ozone (O_3, Tao et al., 2012), particulate matter with an aerodynamic diameter less than 2.5 μgm^{-3} ($PM_{2.5}$, Cesaroni et al., 2013) and less than 10 μgm^{-3} (PM_{10}, Pirani et al., 2014), and sulphur dioxide (SO_2, Wong et al., 2008); where they are related to mortality and morbidity from many diseases including cardio-respiratory diseases. Particulate matter and nitrogen dioxide are amongst the most hazardous pollutants for population health, especially in already sensitive individuals. Particulate matter comprises small solid and liquid particles that are suspended in the air, which when inhaled into the body can travel deep inside the lungs. This makes it an important pollutant and more damaging compared to other air pollutants. Prolonged exposure increases the risks of cardiovascular and respiratory disease, as well as lung cancer. NO_2, on the other hand, is a gaseous pollutant that is caused predominately by traffic and can cause significant inflammation of the airways when exposed to high concentrations. This thesis will focus primarily on NO_2 as data for it are more widely available meaning it is a more useful indicator of air pollution levels compared to the other commonly-used measures, such as particulate matter, for which data are severely sparse across West Central Scotland. There have been many studies investigating the effects of NO_2 on ill health, and a recent study in Scotland found a substantial association with respiratory hospital admissions (Huang et al., 2015).

NO_2 concentrations currently exceed the UK wide objective annual mean concentration of 40 μgm^{-3} set by EU legislation. These guidelines are set to protect the public, but more importantly to protect susceptible members of the population, such as those with asthma who react to lower levels of air pollutants compared to non-asthmatics. In Glasgow, NO_2 concentrations are predicted to exceed these targets until 2020 (Department for the Environment, Food and Rural Affairs, 2015). While the health risk of air pollution to any one person may be small, the risk is substantial in public health terms since a whole population of people will be exposed and thus there is strain on local governments to try and mitigate this (Pope III & Dockery, 2006). Although current air pollution levels are not meeting the regulatory guidelines, air pollution levels in the majority of major western cities have fallen considerably to historically low levels over the last century. Reducing levels of air pollution still remains an area of active research due to the aforementioned WHO estimates.

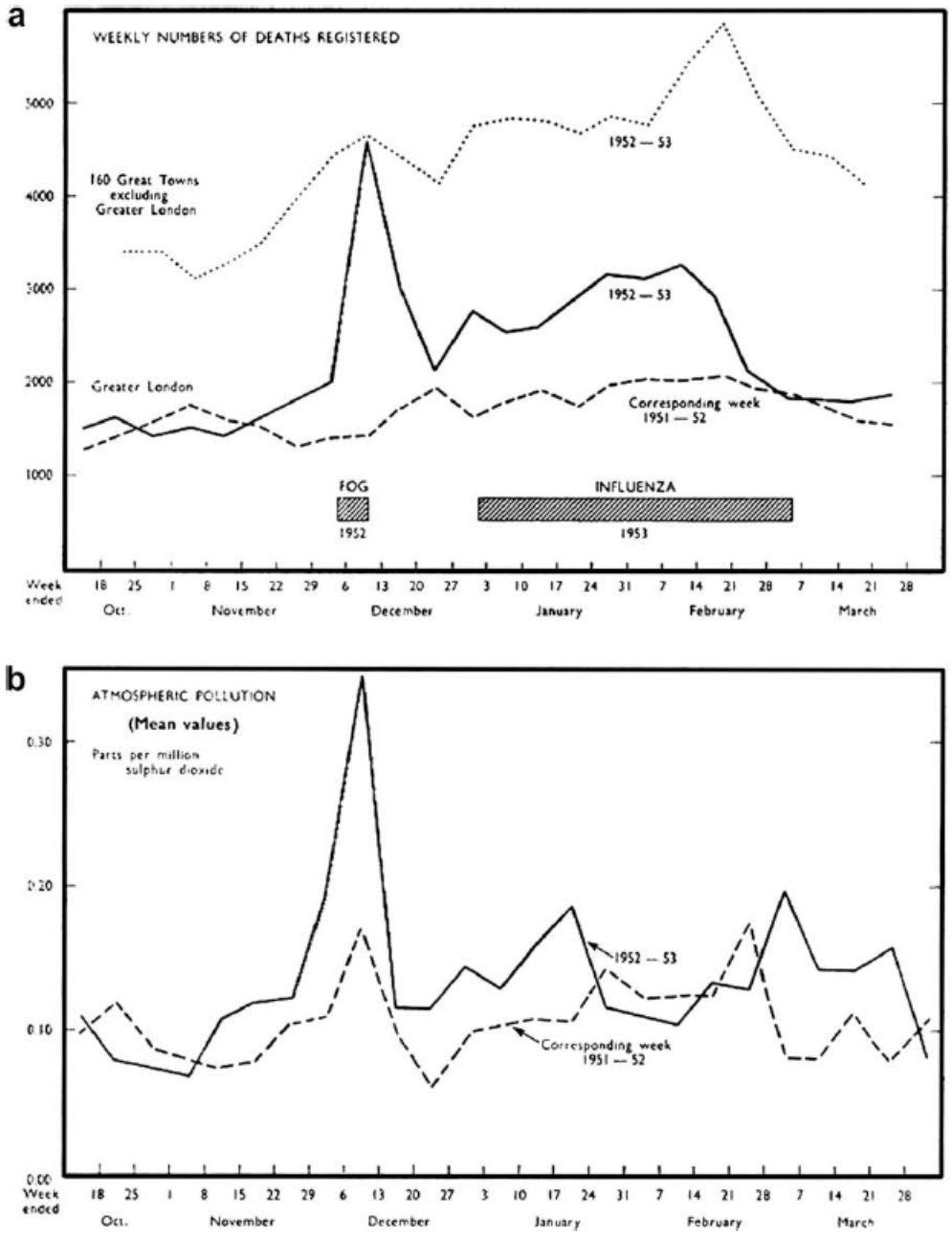

Figure 1.1: *London smog of 1952. Subfigure (a) displays the weekly number of deaths in Greater London compared with those in 140 Great Towns between October 1952 - March 1953. Subfigure (b) displays the mean values of sulphur dioxide compared with 1951-1952. (Anderson, 2009).*

Previous studies have typically focused on the short-term or acute health effects of exposure to air pollution, due to the simplicity, data availability, and quickness to implement and obtain results. These are referred to as time series studies, which estimate the daily effects of air pollution on ill health over short time periods in large urban areas, such as a city or state. This type of study is at the ecological or population level since the health outcome refers to daily aggregated counts of mortality or morbidity, which are then regressed on aggregated daily air pollutant concentrations and other covariates relating to weather, influenza, and the day of the week. Air pollutant concentrations are routinely measured from a number of fixed air pollution monitors that are located throughout the study region, but are typically aggregated to form one overall level of pollution at each time point.

Schwartz & Marcus (1990) were among the first to implement a time series study. This was conducted on data from the Greater London area between 1958 and 1972 and reported a significant association of air pollution with mortality. While each individual time series study is important, there is wide variation in the statistical methods used to explore the relationship between air pollution and health. Therefore, research teams in more recent times have attempted to mitigate this variation through the implementation of large multi-city studies, which standardises the statistical modelling approach and makes sure each city uses the same air pollutants from approved sources to ensure homogeneity between the analyses in the different cities. Examples of these types of studies in Europe include Air Pollution and Health: A European Approach (APHEA, Dab et al., 1996; Katsouyanni & Schwartz, 1996; Samoli et al., 2006), and the National Morbidity, Mortality and Air Pollution Study (NMMAPS, Huang et al., 2005).

The long-term or chronic health effects of exposure to air pollution relate to exposure over a number of years and are typically estimated using cohort studies. Cohort studies follow a number of people at the individual level, rather than at the ecological level, over a specified time frame, which can last a large number of years. Some previous notable cohort studies include the Six Cities study by Dockery et al. (1993) and Laden et al. (2000); the American Cancer Society study by Pope III et al. (2002); the multicentre ESCAPE project in Europe by Beelen et al. (2014); and the Netherlands cohort study by Hoek et al. (2002). However, cohort studies are not always feasible due to the high cost and long implementation, since they require a long follow-up period for every member in the cohort. Therefore, small area spatial ecological studies are instead used to estimate the long-term health effects of air pollution since they are much quicker and easier to implement due to the data being routinely available with no cohort of people to follow up. Haining et al. (2010); Jerrett et al. (2005b); Lawson et al. (2012); Lee et al. (2009); Lee & Sarran (2015); Maheswaran et al. (2005a); Rushworth et al. (2014) were among those to implement such studies.

Small area spatial ecological studies overcome a number of limitations of general ecological studies (e.g., time series) as the populations under study tend to be more homogeneous with respect to their socio-demographic characteristics. Furthermore, unlike time series studies, where a single estimate for air pollution exposure may be used for an entire city, small area studies are able to capture finer spatial variations in ecological exposure levels, which is important for studying the effects of air pollution on ill health. Spatial ecological studies can be cross-sectional to give a snapshot of a moment in time, for example, using data for one year (Lee & Sarran, 2015), or they can also be longitudinal when daily, monthly or consecutive years of data are considered. Spatial ecological studies analyse populations or groups of people rather than individuals, and while they are important for adding to the body of evidence whether an exposure is associated with an outcome, they are not able to establish whether the exposure caused the outcome.

Spatial ecological studies estimate the relationship between air pollution and ill health by modelling geographical contrasts in air pollution and disease risk across areal units determined by administrative boundaries, such as census tracts or postcodes. The disease data comprise counts of the numbers of disease cases within each areal unit, where typically Poisson log-linear models are used to model this relationship. These models also take known confounders, such as socio-economic deprivation into account. However, the spatial patterns in the disease data are never fully accounted for by the covariates and typically contain residual spatial autocorrelation. The leftover spatial patterning can be due to numerous factors, such as unmeasured confounding when an important spatially correlated variable is not included as a covariate in the model. To explain this leftover spatial autocorrelation, the linear predictor containing the covariates includes an additional variable known as a random effect. This set of random effects are typically modelled by a conditional autoregressive (CAR, Lee, 2011) specification, which is a type of Markov random field. The spatial autocorrelation between the random effects is determined by a neighbourhood matrix, where the most common approach is to specify 'neighbours' as areas sharing a common border. This matrix holds information on whether the random effects are partially correlated or not and this correlation can be modelled by a number of CAR specifications, such as the intrinsic model (Besag et al., 1991), and the Leroux model (Leroux et al., 1999). Further details on how these spatial models differ in their specification can be found in Chapter 3.

A major issue in not only spatial ecological studies, but all studies quantifying the health impact of air pollution, is ensuring the air pollutant concentrations used are spatially representative and accurate. Typically, data on air pollutant concentrations are available directly from air pollution monitors that are located throughout the study region, or are available as estimated concentrations on a regular grid that are output from a mathematical atmospheric dispersion model. As monitoring networks are ex-

tremely sparse, using data collected by monitoring stations can be problematic for adequately assessing exposure to air pollution for the whole population under study, as it is not possible to provide every areal unit with an air pollutant value. Therefore, modelled concentrations are used instead as they are estimated at a fine scale resolution and provide complete spatial coverage of the study region. However, they are known to contain biases and contain no measure of uncertainty (Berrocal et al., 2010b).

Another problem with measuring air pollution is that measured and modelled values of air pollution tend to be on different spatial scales, while the disease data are also at another spatial resolution. In the statistical literature this is known as the *change-of-support* problem (Gelfand et al., 2001; Gotway & Young, 2002), and has received great attention in the past decade. While studies can either use measured data on their own (Elliott et al., 2007), or the modelled data on their own (Lee et al., 2009), new methods are now being developed which combine both sets of air pollution data by scaling them to a specific spatial resolution (Berrocal et al., 2010b), or by fusing the two sets of data together (Fuentes & Raftery, 2005).

1.2 Measuring deprivation

How healthy a person is, according to the WHO, is related to numerous social factors, including level of education, whether in employment, level of income, gender and ethnicity. This leads to wide disparities in the health status of different social groups, where a lower socio-economic position implies a higher risk of poor health.

There has been evidence to suggest that deprivation influences the association between air pollution and ill health, and is therefore an important confounding variable to be considered in epidemiological studies. However, studies in Scotland have not been consistent in finding an association between air pollution and ill health, when deprivation is included in the statistical model. This may be due to the multi-factorial nature of deprivation meaning it cannot be fully captured by one or two specific indicators, such as education or employment.

Some studies try and capture deprivation as a whole by utilising an overall measure, such as the Carstairs Score (Carstairs, 2001; Elliott et al., 2007), the Townsend Index (Haining et al., 2010; Maheswaran et al., 2005a, 2006, 2012; Townsend et al., 1988; Walters et al., 1995), or the English Indices of Deprivation (Bennett et al., 2014; Maheswaran et al., 2012; Tonne et al., 2008, 2010). Lastly, studies that do account for numerous indicators (Goodman et al., 2011; Jerrett et al., 2005b; Lee & Mitchell, 2014; Naess et al., 2007) often conclude their research by either selecting one model based on minimising a goodness-of-fit criteria, or presenting the results of all models.

While these are simple approaches at attempting to find the best overall model or to give an overall view of which deprivation indicators are important, they discard all the information present from other models when selecting one, or fail to account for the uncertainty in the choice of deprivation indicator when estimating the pollutant-health effect. Furthermore, different deprivation indicators can result in a wide variation of effect sizes, which are generally not explicitly discussed.

1.3 Aims

Precise measures of air pollution are an important requirement for adequately quantifying the effects of air pollution on ill health. The primary aim of this thesis is to utilise multiple air pollution data sources for extending downscaling and fusion methods in order to increase the accuracy of predicted concentrations for use in future health studies. As estimates of the effect of air pollution on health may be influenced by the choice of air pollutant concentrations and/or deprivation indicator used, this thesis will also investigate the sensitivity of the pollutant-health relationship to these choices and suggest ways of accounting for uncertainty.

1.3.1 Contribution to literature

This thesis will contribute to the body of evidence on air pollution and ill health in Scotland. It will estimate the association between air pollutant concentrations, specifically NO_2, and ill health across West Central Scotland, where ill health is measured in terms of both mortality and hospital admissions as a result of cardio-respiratory disease.

This study will be conducted within an ecological small area framework, which will relate air pollutant concentrations and other covariates, such as socio-economic deprivation to ill health. However, as air pollutant concentrations are generally measured in more urban environments they tend to highlight peak pollutant levels. Furthermore, measurements are usually not evenly distributed across the study area, especially in more rural environments, and are therefore not spatially dense. For a small area study to be feasible, it is crucial to have air pollutant measurements at the same spatial resolution at which the disease data are available. Otherwise, the study lacks statistical power to obtain accurate results. This thesis will develop new spatial methodology for estimating fine-scale air pollutant concentrations that ensures availability of concentrations at the area level that can be aligned with the disease data. This thesis will then relate these new estimated air pollutant concentrations to ill health, while taking into account other covariates, such as socio-economic deprivation. Furthermore, air pollution levels in Scotland are relatively low, so it is important to investigate whether any relationship holds at low pollutant levels, which will help inform any future policy decisions regarding improving air quality.

1.4 Overview of thesis

The remainder of this thesis is split into seven chapters. Chapter 2 provides an overview of the existing statistical methods that form the basis of the methodology utilised in this thesis. It discusses both maximum likelihood methods and Bayesian inference, since analyses are conducted within a Bayesian framework. This chapter delves into generalised linear modelling, while also providing background information on spatial statistics in terms of both geostatistics and areal unit statistics. The geostatistics section provides background methodology to the pollutant model developed in Chapter 4, while the areal unit section provides background methodology for the pollutant-health modelling conducted in Chapters 5 and 6.

Chapter 3 provides a detailed literature review on air pollution and health studies. This chapter discusses further the types of studies, such as time series and cohort studies, used to quantify the impact of air pollution on ill health, with the main focus being on ecological areal unit studies. In addition, Chapter 3 discusses the study design, frequency of disease, and data used throughout remaining chapters. This chapter includes a review of evidence from around the world and how it relates to research within Scotland. The problem of ecological bias is discussed since it plays an important role in spatial ecological studies as one cannot assume association at the ecological level holds at the individual level. This chapter ends with a discussion surrounding the ways in which studies estimate exposure to air pollution at the ecological level, by outlining the difficulties in estimating exposure from a sparse monitoring network. Following this, there is discussion around the approaches employed to enhance spatial prediction of air pollutant concentrations.

Chapter 4 considers the difficulties in estimating air pollutant concentrations that are at the correct spatial resolution to be able to be used alongside the disease data. This chapter discusses the types of air pollutant data that are available in Scotland and how they can be combined to produced accurate fine scale estimated concentrations. The statistical methodology developed to achieve this is based on methods already developed in the statistical literature, but extends them to make use of additional inputs that are available in Scotland, and how they can be combined to produce accurate fine scale estimated concentrations. Several models looking into how different covariates and frameworks (maximum likelihood versus Bayesian) can alter the prediction performance are compared, while utilising procedures to determine the best overall model (in terms of goodness-of-fit criteria). The final model is then used to predict fine scale air pollutant concentrations across West Central Scotland, which can then be used to investigate whether they have a detrimental effect on ill health.

Chapters 5 and 6 relate the new fine scale predicted air pollutant concentrations to cardio-respiratory ill health in West Central Scotland. Both chapters quantify the impact of air pollution on human health, with results from Chapter 5 advocating the statistical model, the set of air pollutant concentrations, and the indicators of socio-economic deprivation to be used in Chapter 6. Chapter 5 utilises a relatively underused type of statistical analysis used in the air pollution and health literature, and aims to combine the results from multiple models into a single overall estimate that takes model uncertainty into account. This stems from studies only selecting the best model that minimises some goodness-of-fit criteria and ignoring information available from other models. Therefore, this chapter develops Bayesian model averaging for use in an air pollution and health context, which is a method to statistically combine results from numerous models into an overall effect size for the association between air pollutant concentrations and ill health. This methodology also helps to establish which of the factors (in terms of the statistical framework employed, set of air pollutant concentrations and choice of socio-economic deprivation utilised) in a model have the greatest contribution to the overall effect size. This will help inform future studies of the factors that are the most important. Chapter 6 utilised a similar approach since deprivation still plays an important confounding role in air pollution and health studies, and made use of the predicted air pollutant concentrations, but this chapter differs from the previous chapter in that the main focus is on disease incidence (number of new cases of a disease within a population) rather than mortality. Incidence was the outcome in this chapter rather than mortality so that there could be a chance of greater understanding of the burden of air pollution on populations that are considered to be healthy with no known pre-existing history of cardio-respiratory disease.

Finally, Chapter 7 presents the key findings from this thesis and outlines the inherent limitations when conducting research of this kind, while proposing ideas for future research. Suggestions of ways in which to engage policy on improving overall air quality are provided.

Chapter 2

Review of Statistical methods

2.1 Introduction

This chapter forms the basis for the statistical techniques used throughout this thesis. Likelihood-based methods of analysis are the predominant approach in many early and current air pollution and health studies (see Larrieu et al., 2007; Prescott et al., 1998; Willocks et al., 2012), and form the basis of more complex techniques. These methods are described in Section 2.2, which details regression methods, such as Poisson regression and quasi-Poisson regression. As time has progressed, data and statistical models have increased in size and complexity, thus resulting in the Bayesian approach becoming the main framework in which analysis is conducted. Therefore, in this thesis, statistical models are implemented within a Bayesian setting. These are introduced in Section 2.3. This section discusses principles of Bayesian analysis, beginning with Bayes Theorem, followed by a discussion of prior distributions and finally how to implement it.

While a Bayesian framework is adopted in this thesis, the relationship between air pollution and ill health is analysed using spatial statistics, which is described in Section 2.4. This section is further divided into two parts: geostatistics and areal unit statistics. Geostatistics, as described in Section 2.4.1, forms the basis for the geostatistical fusion model implemented in Chapter 4. Areal unit data, as described in Section 2.4.2, serves as the basis for the pollution-health modelling implemented in Chapters 5 and 6, but also relates to the different regression methods given in Section 2.2. Finally, Section 2.5 details the methods used to perform direct and indirect standardisation on the health count data to be used in the pollution-health modelling.

2.2 Generalised linear models

Regression is a statistical technique used to determine if there are linear relationships between multiple variables and answers questions such as 'is there a relationship be-

tween air pollution and health?', 'how strong is the relationship between air pollution and health?', and 'how accurately can pollution be predicted in the future?' (Dalgaard, 2008; Faraway, 2004; James et al., 2013). Linear regression is the most straightforward approach for predicting a quantitative response or dependent variable represented by a vector of observed data comprising m observations, $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$, while using the information from one or more predictors (independent variable, explanatory variable, covariate) $x_1, \ldots, x_p$, where p denotes the number of covariates. When $p = 1$ there is only one covariate and regression is called simple linear regression, whereas multiple regression is when $p > 1$. Multiple regression assumes there is an approximately linear relationship between the covariates and response and the model is given by

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \quad i = 1, \ldots, m, \\ &= \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \end{aligned} \tag{2.1}$$

where the errors ϵ_i are assumed to be independent and identically distributed as $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, $\mathbf{x}_i^\top = (x_{0i}, x_{1i}, \ldots, x_{pi})$ and $x_{0i} = 1$ for the intercept term. In vector form the regression model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\mathbf{X}$ is the design matrix of covariates, i.e., $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_m)^\top$. The $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ parameters quantify the association between each covariate and the response and are interpreted as the increase per unit change in each covariate, holding all other covariates as fixed. The parameters $\boldsymbol{\beta}, \sigma^2$ are unknown quantities and are estimated using the method of Maximum Likelihood (ML), where the estimate for the regression coefficients $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ and the estimate for σ^2 is given by $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(m - p)$.

However, in linear regression, the residuals $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ must be continuous and normally distributed. Therefore, it is not an appropriate modelling framework for discrete data, such as binary outcomes, or for count data where the counts can be heavily skewed. Instead, generalised linear models (GLMs, McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972) should be used. These are an extension of the linear modelling framework outlined above, characterised by their response distribution, p, and a link function, G, which describes how the mean, μ, is transformed onto a scale related to the linear predictor. In general terms, the linear predictor of a regression model is given by $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and the link function G (it can be noted that McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972 actually denote G^{-1} as the link function) is

the function that relates the mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}|\mathbf{X}]$ and the linear predictor $\boldsymbol{\eta}$ by

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = G(\mathbf{X}\boldsymbol{\beta}) \Leftrightarrow \boldsymbol{\mu} = G(\boldsymbol{\eta}). \tag{2.3}$$

The linear model is a special case of the GLM, where the link function is simply the identity function. Common link functions include log and square root transformations, whereas in a logistic regression analysis where the response is binary coming from a Binomial distribution, the link function is $\text{logit}(l_i) = \log(l_i/(1-l_i)) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where l is the probability of an event occurring.

As the disease data in this thesis are based on counts of mortality and hospital admissions due to cardio-respiratory disease, linear regression may produce biased results (Coxe et al., 2009). Instead, Poisson regression can be used to appropriately model count data. This is a special case of a generalised linear model and is discussed below in Section 2.2.1.

2.2.1 Poisson and quasi-Poisson regression

A count variable is one that can only take positive integer values such as $(0, 1, 2, \dots)$ and reflects the number of occurrences of an event, such as the number of people in a specific area that have died, in a specified time frame. Count variables can only be positive (or zero) since an event cannot occur a negative amount of times. Using count data in ordinary least squares (OLS), which is the optimisation method used in linear regression, may violate some of the assumptions that are placed on the error structure of the model. The errors of a model are given by the residuals. A residual is the difference between the observed data Y_i and the predicted data $\hat{Y}_i$ and is given by $\hat{e}_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$. The assumptions that are placed on the error structure are that of normality, constant variance (homoscedasticity) and independence. Count data can violate these assumptions when the variance increases with the mean of the data (heteroscedasticity) and by displaying skewness in its distribution.

In order to overcome these assumption violations, Poisson regression can be utilised in which the Poisson distribution represents the distribution of the errors. As stated above, Poisson regression falls into the class of generalised linear models where the outcome is transformed to linearise the relationship between the response and the covariate/s. This transformation is performed through the link function, in which the natural log is the link function used in Poisson regression. The Poisson distribution is important for modelling count data since it is a discrete distribution that only takes positive integers and the probability mass function for the Poisson distribution is given

by

$$p(Y = y|\mu) = \frac{\mu^y}{y!}\exp(-\mu), \quad y = 0, 1, \ldots, \tag{2.4}$$

where μ is both the mean and variance of the distribution, i.e., $\mathbb{E}[Y] = \text{Var}[Y] = \mu$. The Poisson regression model for response Y_i and covariates $\mathbf{x}_i^\top = (x_{0i}, x_{1i}, \ldots, x_{pi})$ is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i), \quad i = 1, \ldots, m, \\ \log(\mu_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \end{aligned} \tag{2.5}$$

where μ_i is the expected count given specific values of the covariates. As the transformed outcome is no longer on the same scale as the original outcome, to find the value of μ_i for specified values of the covariates, the exponential of the linear predictor is taken, i.e., $\exp(\mathbf{x}_i^\top \boldsymbol{\beta})$.

One of the main issues with health count data is that they typically exhibit greater variation compared to the mean. This poses an issue for Poisson regression since the Poisson distribution has an equal mean and variance. When the variance is greater than the mean it is known as overdispersion (underdispersion when the variance is less than the mean and equidispersion when the mean and variance are equal). One of the most common ways of dealing with overdispersed health count data is to use quasi-Poisson regression (Wedderburn, 1974). Quasi-Poisson regression is similar to Poisson regression except it has two parameters μ and ϕ, where ϕ is the parameter that accounts for any overdispersion present in the data. The quasi-Poisson model is characterised by its mean as $\mathbb{E}[Y_i] = \mu_i$, and variance as $\text{Var}[Y_i] = \phi\mu_i$, where the variance is assumed to be a linear function of the mean (Ver Hoef & Boveng, 2007). Quasi-Poisson models are estimated by maximum likelihood (ML) using the method of iteratively weighted least squares (IWLS), which is outlined below.

2.2.1.1 IWLS

In the GLM framework, the distributions used to characterise the observed data such as Gaussian, Binomial, Exponential and Poisson, are members of the exponential family of distributions, where the probability distributions can be written in the form

$$p(Y = y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \tag{2.6}$$

for some random variable Y, where ϕ is the dispersion parameter, θ is the parameter of interest (canonical parameter), and $a(\bullet), b(\bullet)$ and $c(\bullet)$ are some specific functions. For example, suppose Y is normally (Gaussian) distributed, i.e., $Y \sim \text{N}(\mu, \sigma^2)$, then

the distribution can be written as a member of the exponential family as

$$\begin{aligned} p(Y=y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \\ &= \exp\left(y\frac{\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}-\log(\sqrt{2\pi\sigma^2})\right), \end{aligned} \tag{2.7}$$

where $a(\phi)=\sigma^2, b(\theta)=\mu^2/2$, $c(y,\phi)=-y^2/2\sigma^2-\log(\sqrt{2\pi\sigma^2})$, $\phi=\sigma$ and $\theta=\mu$.

The method of estimation for GLMs is maximum likelihood, where for the vector of regression coefficients $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^\top$ the likelihood is maximised with respect to $\boldsymbol{\beta}$. Denote the vector of observations as $\mathbf{Y}=(Y_1,\ldots,Y_m)^\top$, and denote the vector of expectations as $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_m)^\top$. Given the link function, $\theta_i=\mathbf{x}_i^\top\boldsymbol{\beta}=G(\mu_i)$. Then the log-likelihood of $\mathbf{Y}$ is

$$l(\mathbf{Y},\boldsymbol{\theta},\phi)=\sum_{i=1}^{m} l(Y_i,\theta_i,\phi), \tag{2.8}$$

where $\theta_i=\theta(\eta_i)=\theta(\mathbf{x}_i^\top\boldsymbol{\beta})$, and $l(\bullet)$ on the right hand side of (2.8) denotes the individual log-likelihood for each observation i. In the situation of quasi-likelihood where the distribution of $\mathbf{Y}$ is unknown, the log-likelihood is replaced by the first two moments of the unknown distribution, where it is assumed that $\mathbb{E}[\mathbf{Y}]=\boldsymbol{\mu}$ and $\text{Var}[\mathbf{Y}]=a(\phi)V(\mu)$. The quasi-likelihood (Nelder & Wedderburn, 1972) is then defined by

$$l(\mathbf{Y},\theta,\phi)=\frac{1}{a(\phi)}\int_{\mu(\theta)}^{y}\frac{s-y}{V(s)}ds. \tag{2.9}$$

When $\mathbf{Y}$ is from the exponential family of distributions, then the derivatives of $l(y,\theta,\phi)=\log(p(y|\theta,\phi))$ and the quasi-likelihood coincide.

If the log-likelihood in equation (2.8) is replaced with with the general form of the exponential family of distributions given in equation (2.6), then

$$l(\mathbf{Y},\boldsymbol{\theta},\phi)=\sum_{i=1}^{m}\left[\frac{Y_i\theta_i-b(\theta_i)}{a(\phi)}-c(Y_i,\phi)\right]. \tag{2.10}$$

Since the terms $a(\phi)$ and $c(Y_i,\phi)$ will not have an influence on the maximisation, it is sufficient to consider the log-likelihood to be simplified to

$$\tilde{l}(\mathbf{Y},\boldsymbol{\theta})=\sum_{i=1}^{m}(Y_i\theta_i-b(\theta_i)). \tag{2.11}$$

Then to maximise this log-likelihood with respect to $\boldsymbol{\beta}$, the derivative is taken which

yields

$$\frac{\partial}{\partial \boldsymbol{\beta}} \tilde{l}(\mathbf{Y}, \boldsymbol{\theta}) = \sum_{i=1}^{m} (Y_i - b'(\theta_i)) \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i, \tag{2.12}$$

which is solved by setting equal to zero and rearranging for $\boldsymbol{\beta}$. However, this would require solving a set of nonlinear system of equations, which requires an iterative procedure to be solved. The Newton-Raphson algorithm is an iterative procedure for solving such systems of equations. First of all, denote the Hessian of the log-likelihood to be $H(\boldsymbol{\beta})$, which is the matrix of second derivatives for all elements of $\boldsymbol{\beta}$. One iteration of the Newton-Raphson algorithm for $\boldsymbol{\beta}$, where $\mathcal{D}(\boldsymbol{\beta})$ denotes equation (2.12), is

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (H(\hat{\boldsymbol{\beta}}^{(t)}))^{-1} \mathcal{D}(\hat{\boldsymbol{\beta}})^{(t)}. \tag{2.13}$$

The Fisher scoring algorithm is a variation of the Newton-Raphson method, which replaces the Hessian matrix by its expectation as

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \mathbb{E}[H(\hat{\boldsymbol{\beta}}^{(t)})]^{-1} \mathcal{D}(\hat{\boldsymbol{\beta}})^{(t)}. \tag{2.14}$$

Recall that $\theta_i = \mu_i = G(\mathbf{x}_i^\top \boldsymbol{\beta}) = b'(\theta_i), \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, so $b'(\theta_i) = G(\eta_i)$. Then the first and second derivatives of θ_i are

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i &= \frac{G'(\eta_i)}{V(\mu_i)} \mathbf{x}_i, \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \theta_i &= \frac{G''(\eta_i) V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^\top. \end{aligned} \tag{2.15}$$

Then, the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ given by $\mathcal{D}(\boldsymbol{\beta})$ can be expressed as

$$\mathcal{D}(\boldsymbol{\beta}) = \sum_{i=1}^{m} (Y_i - \mu_i) \frac{G'(\eta_i)}{V(\mu_i)} \mathbf{x}_i. \tag{2.16}$$

And the Hessian matrix can be expressed as

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left[- b''(\theta_i) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \right) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \right)^\top - (Y_i - b'(\theta_i)) \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \theta_i \right] \tag{2.17}$$

$$= \sum_{i=1}^{m} \left[\frac{G'(\eta_i)^2}{V(\mu_i)} - (Y_i - \mu_i) \frac{G''(\eta_i) V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \right] \mathbf{x}_i \mathbf{x}_i^\top \tag{2.18}$$

In the Fisher Scoring algorithm in equation (2.14), the expectation of the Hessian matrix can be replaced by

$$\mathbb{E}[H(\boldsymbol{\beta})] = \sum_{i=1}^{m} \left(\frac{G'(\eta_i)^2}{V(\mu_i)} \right) \mathbf{x}_i \mathbf{x}_i^\top, \tag{2.19}$$

since $\mathbb{E}[Y_i] = \mu_i$. Furthermore, define the weight matrix for the Fisher scoring algorithm as

$$\mathbf{W} = \text{diag}\left(\frac{G'(\eta_1)^2}{V(\mu_1)}, \ldots, \frac{G'(\eta_m)^2}{V(\mu_m)}\right), \tag{2.20}$$

and define

$$\tilde{\mathbf{Y}} = \left(\frac{Y_1 - \mu_1}{G'(\eta_1)}, \ldots, \frac{Y_m - \mu_m}{G'(\eta_m)}\right). \tag{2.21}$$

One iteration for $\boldsymbol{\beta}$ can then be expressed as

$$\begin{aligned} \boldsymbol{\beta}^{(t)} &= \boldsymbol{\beta}^{(t-1)} + (\mathbf{x}_i^\top \mathbf{W} \mathbf{x}_i)^{-1} \mathbf{x}_i^\top \mathbf{W} \tilde{\mathbf{Y}}, \\ &= (\mathbf{x}_i^\top \mathbf{W} \mathbf{x}_i)^{-1} \mathbf{x}_i^\top \mathbf{W} \mathbf{Z}, \end{aligned} \tag{2.22}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_m)^\top$ denotes the vector of adjusted dependent variables, i.e., $z_i = \mathbf{x}_i^\top \boldsymbol{\beta}^{(t-1)} + (Y_i - \mu_i)(G'(\eta_i))^{-1}$. The iteration stops when the parameter estimate or log-likelihood no longer changes significantly. Then, the parameter estimates for regression parameters are denoted by $\hat{\boldsymbol{\beta}}$. At each step of the iteration, weighted least squares is performed on the adjusted responses z_i on $\mathbf{x}_i$. For normal linear regression, iteration is not necessary because the link function is the identity so $G' = 1$ and $\mu_i = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. The second derivative of θ_k is zero when the link is canonical, for example, the log-link for count data. An estimate for the dispersion parameter ϕ can be obtained from

$$\hat{a}(\phi) = \frac{1}{m} \sum_{i=1}^{m} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{2.23}$$

2.3 Bayesian modelling

The foundation of Bayesian inference is Bayes theorem, which was introduced by English statistician and Reverend, Thomas Bayes (Bayes, 1764). Firstly, a joint probability distribution for the vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and observed data $\mathbf{Y}$ can be written as a product of the prior distributions $p(\boldsymbol{\theta})$ and the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ as

$$p(\boldsymbol{\theta}, \mathbf{Y}) = p(\boldsymbol{\theta}) p(\mathbf{Y}|\boldsymbol{\theta}). \tag{2.24}$$

Observing data changes the information about a parameter according to

$$p(\boldsymbol{\theta}) \longrightarrow p(\boldsymbol{\theta}|\mathbf{Y}). \tag{2.25}$$

In other words, prior information via $p(\boldsymbol{\theta})$ relates to the posterior $p(\boldsymbol{\theta}|\mathbf{Y})$, which is the probability of the parameters conditioned on the observed data. The way in which

prior information relates to the posterior is through Bayes Theorem

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) &= \frac{p(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \\ &= \frac{p(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto p(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta}), \end{aligned} \tag{2.26}$$

and is equal to the prior density of unknown parameters $\boldsymbol{\theta}$ times the likelihood of the data $p(\mathbf{Y}|\boldsymbol{\theta})$, divided by a normalising constant $p(\mathbf{Y})$. This can be simplified to be proportional to the prior times the likelihood. The likelihood, $p(\mathbf{Y}|\boldsymbol{\theta})$, is the likelihood of the observed data $\mathbf{Y}$ under a probability model, and the prior is the distribution of knowledge about $\boldsymbol{\theta}$ (set of parameters) before any data are observed. The posterior therefore reflects uncertainty in the parameters after taking the prior information and data into account. Typically, the central value, such as the mean or median, of the parameter's posterior distribution is taken to be its point estimate, where it provides probabilistic statements about the parameter. A $c\%$ credible interval quantifies the uncertainty surrounding the point estimate and estimates that the parameter will lie within a specific interval with probability $\frac{c}{100}$.

The choice of prior distribution is subjective and is one of the criticisms of a Bayesian analysis, but prior information can be based on information from previous studies, expert intuition, or can be chosen to represent prior ignorance or on a basis of convenience. It is common practice to use prior distributions that are close to flat, encouraging a negligible impact on the posterior, which is driven mostly from the data (Gelman et al., 2003). Thus, at this particular point, a frequentist model (using a ML approach) and a Bayesian model should have very similar results. A hierarchical Bayesian framework is typically used, in which the parameters of the prior distribution for each of the parameters is estimated using a further set of probability distributions in terms of further parameters $\boldsymbol{\phi}$ known as hyperparameters, which are given by

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}), \tag{2.27}$$

where $p(\boldsymbol{\phi})$ is the set of hyperpriors and affects $\mathbf{Y}$ only through $\boldsymbol{\theta}$. The model is hierarchical due to its inherent structure, since there are hyperparameters, $\boldsymbol{\phi}$, with distribution $p(\boldsymbol{\phi})$ that governs the prior distributions, $p(\boldsymbol{\theta}|\boldsymbol{\phi})$, which in turn are combined with the data $p(\mathbf{Y}|\boldsymbol{\theta})$ to produce the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$.

2.3.1 Choice of prior distribution

The prior distribution for a model parameter conveys all available information that is known, before observing any data. Therefore, it is an important consideration when preforming Bayesian inference as the choice of prior will influence the posterior distribution. One can choose a univariate prior for each of the parameters, such as $p(\boldsymbol{\theta}) = \prod_{i=1}^{d} p(\theta_i)$, or combine information for all individual model parameters into a single multivariate prior. Typically, such priors are used in combination as is the case in this thesis.

There are many types of prior distributions that can be utilised, and each type of prior has its own merits. The type of prior chosen depends on the information or belief to be conveyed in the model. A conjugate prior (Raïffa & Schlaifer, 1961) is one that results in a posterior density of the same parametric distribution as the prior. For example, a Gamma prior combined with a Poisson likelihood will yield a Gamma distribution for the posterior. The likelihood for a Poisson distribution is given as

$$\begin{aligned} p(Y = y|\mu) &= \frac{\mu^y}{y!}\exp(-\mu) \\ &\propto \mu^y \exp(-\mu), \end{aligned} \tag{2.28}$$

and the Gamma distribution for the prior of μ is given as

$$\begin{aligned} p(\mu|\alpha, \beta) &= \frac{\beta^\alpha \mu^{\alpha-1}\exp(-\beta\mu)}{\Gamma(\alpha)}, \\ &\propto \mu^{\alpha-1}\exp(-\beta\mu). \end{aligned} \tag{2.29}$$

Then, to obtain the posterior distribution for μ, the likelihood is multiplied by the prior; however, the posterior is proportional to these elements. Therefore, in the likelihood and prior distributions, only terms that are related to μ are kept as any terms not containing μ are normalising constants. Therefore, combining the likelihood and the prior yields a posterior density of

$$\begin{aligned} p(\mu|y) &\propto \mu^{y+\alpha-1}\exp(-\mu(\beta+1)), \\ &= \text{Gamma}(\mu|y+\alpha, \beta+1). \end{aligned} \tag{2.30}$$

A conjugate prior is used for 'algebraic convenience' as the resulting posterior is of the same parametric form as the prior and is therefore easy to understand and evaluate in closed-form.

Informative priors are such that the prior and the likelihood both have an influence

on the resulting posterior density. These types of priors must be handled with care, but they demonstrate the power of Bayesian methods as information on previous studies, expert opinion or experience can be combined with current knowledge in a natural way.

Flat or noninformative prior specifications are used so that the resulting posterior distribution is driven by the data, thus reducing the amount of subjective belief to be incorporated into the model. A typical flat prior is the unbounded uniform distribution from negative infinity to positive infinity for a parameter θ on the real line. This does indeed allow the data to determine the posterior density, but the resulting posterior is improper as it cannot integrate to one and is therefore not a valid probability distribution. However, if the uniform prior is on the interval $[0, 1]$ for a parameter characterising a proportion, then the resulting posterior distribution will be proper.

An alternative is the class of weakly informative priors, which are useful when there is a lack of knowledge surrounding a parameter as little or no information is known. These priors attempt to be noninformative, but are still proper priors that integrate to one as they are not fully flat like the uniform prior. Furthermore, weakly informative priors are purposely constructed so that the information contained is weaker than any prior knowledge that is actually available (Gelman, 2006) and have a negligible effect on the posterior, which allows the observed data to determine the posterior density. An example of a weakly informative prior is when assigning a multivariate Gaussian prior with mean zero and a large diagonal variance matrix (for example, setting the diagonal elements to 1000) to the set of regression parameters for p covariates in a regression model.

2.3.2 Inference

Bayesian inference can be conducted in many ways, but the most common approach to obtaining the posterior distribution is through Markov chain Monte Carlo (McMC) simulation methods. These methods allow sampling from the posterior distribution by constructing a Markov chain, which produces correlated realisations from the posterior after a finite number of iterations. The samples are drawn sequentially where the current sample only depends on the previous sample. This is known as a Markov chain. In probability theory, a Markov chain is a stochastic process comprising M random variables $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)})$, where for any sample t, the distribution of $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ given all previous values only depends on the previous value, $\boldsymbol{\theta}^{(t-1)}$ and d is the number of parameters. This Markov property is fundamental to McMC simulation, but the key property of McMC is that it is an iterative procedure in which the approximate distributions are improved at each step of the algorithm, since they converge to the target (posterior) distribution $p(\boldsymbol{\theta}|\mathbf{Y})$. The simulation has to be run long enough so that the distribution of the current draws are the target

density. The Markov chain is then assessed for convergence as the simulated sequences should come from the target density.

Within McMC methods, there are numerous algorithms for sampling from the posterior distribution such as Gibbs sampling, and the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is the simplest and most general McMC algorithm (Banerjee et al., 2004; Robert & Casella, 2010). This algorithm is used to sample from the posterior $p(\boldsymbol{\theta}|\mathbf{Y})$ when the posterior distribution is not a standard distribution (such as a Gaussian distribution), or when some/all of the full conditional distributions are not standard distributions. Consider a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. The full conditional distribution of θ_i is the distribution of the parameter conditioned on the known information and all other parameters, that is $p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, \mathbf{Y})$.

The Metropolis-Hastings algorithm is an adaptation of the Metropolis algorithm, which uses an acceptance/rejection rule to converge to the specified posterior or full conditional distribution for each parameter. The Metropolis algorithm (Metropolis et al., 1953) is outlined below.

1. Choose starting values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$, for which $p(\boldsymbol{\theta}^{(0)}|\mathbf{Y}) > 0$, from a starting distribution $p(\boldsymbol{\theta}^{(0)})$.

2. At iteration t, for parameters $i = 1, \ldots, d$:

   (a) Draw a proposal value $\theta_i^{(*)}$ from a jumping or proposal distribution, $J_t(\theta_i^{(*)}|\theta_i^{(t-1)})$. The jumping distribution must be symmetric for the Metropolis algorithm such that $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$, for all θ_a, θ_b and t.

   (b) Compute the acceptance ratio (probability),

   $$r = \frac{p(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}, \mathbf{Y})}{p(\theta^{(t-1)}|\boldsymbol{\theta}_{-i}, \mathbf{Y})}. \tag{2.31}$$

   (c) Accept $\theta_i^{(*)}$ as $\theta_i^{(t)}$ with probability $\min(r, 1)$. If $\theta_i^{(*)}$ is not accepted, then $\theta_i^{(t)} = \theta_i^{(t-1)}$.

3. Repeat step 2 M times to get M draws from $p(\boldsymbol{\theta}|\mathbf{Y})$.

The proposal distribution determines where the chain moves to in the next iteration, thus the support of the proposal distribution must contain the support of the posterior or full conditional distribution. Furthermore, it is important to monitor the acceptance rate (proportion of proposal values that are accepted) of the algorithm because, if it is too high, the chain may not be mixing well (i.e., the chain is not moving around the parameter space quickly enough). If the acceptance rate is too low, the algorithm is too inefficient as it is rejecting too many proposal values. For the Metropolis-Hastings algorithm (Hastings, 1970), the jumping distribution needs no longer be symmetric.

To correct for the asymmetry in the jumping distribution, the acceptance ratio in step 2 is replaced with

$$r = \frac{p(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}, \mathbf{Y})/J_t(\theta_i^{(*)}|\theta_i^{(t-1)})}{p(\theta_i^{(t-1)}|\boldsymbol{\theta}_{-i}, \mathbf{Y})/J_t(\theta_i^{(t-1)}|\theta_i^{(*)})}. \tag{2.32}$$

The Gibbs sampler (Geman & Geman, 1984) is a special case of the Metropolis-Hastings algorithm when each full conditional distribution is a known standard distribution. Thus, the Gibbs sampler is useful for conditionally conjugate models, where one can sample from each conditional posterior distribution. The procedure for Gibbs sampling is outlined below.

1. Choose a vector of starting values $\boldsymbol{\theta}^{(0)}$.

2. At iteration t, for parameters $i = 1, \ldots, d$:

   (a) Draw a value $\theta_1^{(t)}$ from the full conditional distribution $p(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, \mathbf{Y})$.

   (b) Draw a value $\theta_2^{(t)}$ from the full conditional distribution $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)}, \mathbf{Y})$, where $\theta_1^{(t-1)}$ has been replaced by its updated value $\theta_1^{(t)}$.

   $\vdots$

   (c) Draw a value $\theta_d^{(t)}$ from the full conditional distribution $p(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{-d}^{(t)}, \mathbf{Y})$, using updated values for $\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{d-1}^{(t)}$.

3. Repeat step 2 until M draws are obtained, with each draw being a vector $\boldsymbol{\theta}^{(t)}$.

Thus, each individual parameter is updated conditional on latest values of the remaining parameters, which are the iteration t values for the parameters already updated and iteration $t-1$ values for the others.

2.3.3 Diagnostics

As said previously, it is important to assess the Markov chain to ensure that it is converging to the target density. In order to diminish the effect of the starting distribution, a specific number of samples from the beginning of the Markov chain are discarded, which is known as the burn-in period. Another consideration is that of thinning the Markov chain by only keeping every kth iteration for each parameter and discarding the rest of the samples. This is performed as a way of breaking the dependence between iterations so that the chosen iterations are not too correlated. This also reduces the memory needed on the computer to save all the samples since only a fraction of

the samples are being kept. However, thinning can increase the variance of the estimates (Gelman et al., 2003), increases the computational time required to obtain all draws, and it has been argued that it is inefficient (Link & Eaton, 2012). The simplest way of assessing convergence of the Markov chain is to study trace plots for each parameter. Trace plots display the samples versus the simulation index (iterations) and show whether the chain has reached its stationary or target distribution. A chain is considered to have reached its stationary or target distribution if the distribution of samples has relatively constant mean and variance, as shown in Figure 2.1. The trace plot on the left is centred around a specific value and the density plot does not display any skewness. A trace plot can also show whether the chain is mixing or not. A chain that is mixing well will jump from one remote region to another in relatively few steps. Furthermore, it is important for the chain to mix and have the ability to explore the parameter space so that a good estimate of the posterior distribution can be provided. This is assessed by examining the acceptance rates for each parameter, where a low acceptance rate indicates the proposal values are not being accepted due to vast exploration of the parameter space beyond the support of the posterior density. In contrast, when too many proposal values are accepted it indicates the chain is not mixing and thus the acceptance rate is too high.

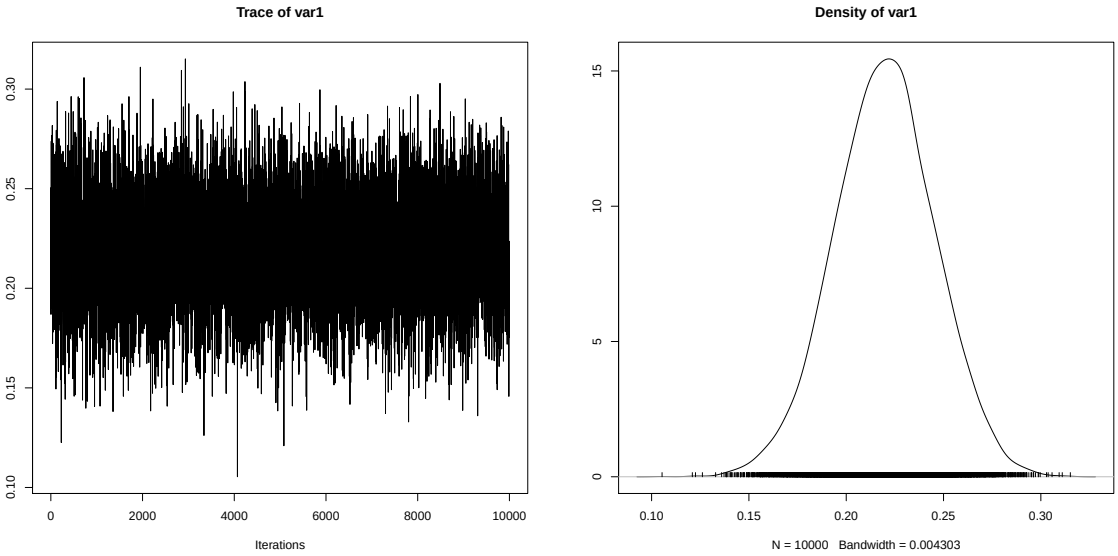

Figure 2.1: *Trace plot for a model parameter is shown on the left and the density of the iterations is shown on the right.*

There are numerous formal diagnostic tests for assessing the convergence of Markov chains, such as Gelman and Rubin diagnostics (Brooks & Gelman, 1998; Gelman & Rubin, 1992), which uses parallel chains to test whether they all converge to the same target distribution; Heidelberger and Welch Diagnostics (Heidelberger & Welch, 1981, 1983), which assesses whether the Markov chain is a weakly stationary process; and Geweke diagnostics (Geweke, 1992), which tests whether the mean estimates have converged by comparing the mean estimate from the beginning of the Markov chain

to the mean estimate from the end of the Markov chain. For the Geweke test, if the comparison of the mean estimates between the beginning and end of the Markov chain are not similar then the chain has failed to converge. The test is a two-sided test based on a z score where a large score value indicates failure of convergence. The mean for the set of M samples for θ_i is given as

$$\hat{\theta}_i = \frac{1}{M}\sum_{t=1}^{M}\theta_i^{(t)}. \tag{2.33}$$

Geweke's diagnostic is then calculated by taking the difference between the mean of $\boldsymbol{\theta}^{(t)}$ based on the first set of m_1 samples given as $\hat{\theta}_1$, and the mean of $\boldsymbol{\theta}^{(t)}$ based on the last set of m_2 samples given by $\hat{\theta}_2$ and dividing by the asymptotic standard error of the difference, where the sample variances $\hat{s}_1, \hat{s}_2$ are computed using spectral densities. If the ratios m_1/M and m_2/M are fixed and $m_1 + m_2 < M$, and the chain is stationary then the following statistic converges to a standard normal distribution as $M \to \infty$

$$Z_M = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\hat{s}_1}{m_1} + \frac{\hat{s}_2}{m_2}}}. \tag{2.34}$$

2.3.4 Model comparison and selection

When multiple statistical models are at play, it is of interest to have a method which allows the researcher to compare models and thus determine which model provides the best fit to the data. It is necessary to utilise a model selection technique that balances out the goodness of fit of a model, as determined by the model likelihood, with its complexity, as determined by its number of parameters. More complex models contain more parameters, which can lead to these models overfitting the data even though the goodness of fit will be high. Therefore, it is important that a model comparison and selection technique balances out these two phenomena. The approaches described below are only viable when two or more models are being compared since they cannot provide any information about the quality of a model in an absolute sense - it is only a measure of the relative quality of the model.

One of the most common approaches to model comparison and selection is the Akaike Information Criterion (AIC; Akaike, 1973) approach, which penalises models that are too complex in terms of overparameterisation by containing a term for the number of model parameters. The AIC is defined as

$$\text{AIC} = -2\log(\hat{L}) + 2q, \tag{2.35}$$

where $\hat{L}$ is the maximum likelihood of the candidate model, and q is the number of model parameters. Given a set of candidate models, the model with the lowest AIC value is deemed the more appropriate model.

Another approach to model comparison and selection is the Bayesian Information Criterion (BIC; Schwarz, 1978), which is similar to AIC in the sense that the model with the lowest value of the BIC is the preferred model of choice. The BIC is defined as

$$\text{BIC} = -2\log(\hat{L}) + q\log(m), \tag{2.36}$$

where m is the number of observations in the model. The difference between the BIC and the AIC is that the BIC penalises the number of model parameters more strongly compared to the AIC since the penalty is via $\log(m)$ rather than by 2.

A more common type of model comparison and selection technique within a Bayesian setting is the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), which again involves a trade-off between the goodness of fit of the model and the model's complexity. The DIC is defined as

$$\text{DIC} = \overline{D} + p_D, \tag{2.37}$$

where $\overline{D} = \mathbb{E}[-2\log(\hat{L})]$ is the posterior mean deviance which measures the goodness of fit of the model, and p_D is the effective number of parameters to assess complexity. Again, given a set of candidate models, the model with the lowest DIC value is the preferred choice. Furthermore, the DIC is similar to the BIC since it penalises models which have extra unnecessary parameters and therefore prefers more parsimonious models. For futher details on these model comparison, selection techniques and compairson of AIC and DIC within a Bayesian context see Gelman et al. (2014).

2.4 Spatial statistics

Spatial statistics is the quantitative analysis and modelling of observed data at different geographical locations. These geographical locations are typically in 2-dimensional space consisting of (x, y) co-ordinates, such as latitude and longitude; however, they can also be in 3 dimensions if considering elevation from the earth. Due to the nature of spatial data, there is no ordering of the observations, unlike in time series data which are naturally ordered in time. Furthermore, independence of the observations cannot be assumed as observations closer together in space are more likely to have similar values and this dependence means that commonly-used statistical methods requiring the assumption of independence are not appropriate.

A geographer named Waldo Tobler stated that '*Everything is related to everything else, but near things are more related than distant things*'. This is known as Tobler's first law of geography (Tobler, 1970), and is the basis for spatial analysis as it requires the spatial dependence between observations to be modelled. Ignoring this dependence hinders the quality of statistical inference.

There are three main types of spatial data, with only the first two utilised in this thesis. These are: geostatistical data, areal unit data and point data. Geostatistical data are data that could potentially be measured at any location within a 2-dimensional region. For example, in a city, air pollution could be observed at infinitely many locations, but in practice there is a limited time frame and budget meaning that data can only be collected at a fixed number of locations. For areal unit data, the study region of interest is stratified into a number of non-overlapping subregions, such as electoral wards and data, such as air pollution concentrations are observed for each of the subregions. For point data, the locations of the observations are themselves data, unlike in geostatistical data, with the number of locations being random rather than specified by the data collector. For example, the location at which a lightening bolt strikes the earth.

The objective of a spatial analysis is different depending on the type of spatial data obtained. For geostatistical data, the goal is to identify and understand the spatial pattern in the data by finding a statistical model to explain the observed spatial dependence between observations. In addition, geostatistics also seeks out to predict the spatial process at unmeasured locations by utilising the data already observed. For areal unit data, one of the goals is to also understand the spatial pattern in the data, but to also utilise ecological regression in order to estimate the effects of a predictor on a response, while taking into account the fact that the residuals of the model will be spatially autocorrelated. And lastly, for point data the goal is to find a statistical model to explain the spatial dependence in the data, while estimating the intensity of the event, i.e., the more points in a region, the higher the intensity.

The remainder of this section discusses the methodology behind geostatistics and areal data in more detail in order to form the basis for the statistical methodology used in this thesis.

2.4.1 Geostatistics

As discussed above, geostatistical data arise when data are collected at a fixed number of locations within a specified study region. A geostatistical process is a stochastic

process where the spatial domain, D, is a fixed subset of the 2-dimensional space

$$\{Z(\mathbf{s})\colon \mathbf{s} \in D \subset \mathbb{R}^2\}, \tag{2.38}$$

where $Z(\mathbf{s})$ is a random variable representing the quantity of interest at spatial location $\mathbf{s}$, and $\mathbb{R}^2$ is a continuous 2-dimensional region of which D is a subset. In reality, data are observed at a finite number of locations m and are denoted by $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))$. The main aim of spatial analysis is to model the spatial dependence in the data. Geostatistical data will be positively correlated, meaning that two observations closer together in space are more likely to have similar values. This correlation is due to the variable of interest being affected by other unmeasured variables that are also spatially correlated. For example, air pollution concentrations are spatially correlated since air pollution is caused by traffic emissions, and two air pollution monitoring stations on the same road will produce similar pollution levels.

2.4.1.1 Properties of the geostatistical process

The mean function or first moment of the stochastic process is defined by

$$\mu_Z(\mathbf{s}) = \mathbb{E}[Z(\mathbf{s})], \quad \forall\, \mathbf{s} \in D. \tag{2.39}$$

The mean function, $\mu_Z(\mathbf{s})$, varies along space and can be interpreted as the expectation at location $\mathbf{s}$, taken over the distribution of all possible values that could have been generated from the stochastic process $\{Z(\mathbf{s})\}$. The stochastic process can either be a continuous random variable or a discrete random variable and when $Z(\mathbf{s})$ is continuous the mean function is given by

$$\mu_Z(\mathbf{s}) = \mathbb{E}[Z(\mathbf{s})] = \int_{-\infty}^{+\infty} z f_{Z(\mathbf{s})}(z)\, dz, \tag{2.40}$$

where $f_{Z(\mathbf{s})}(\cdot)$ is the probability density function (pdf) for stochastic process $Z(\mathbf{s})$. When $Z(\mathbf{s})$ is a discrete random variable with sample space Ω, the mean function is given by

$$\mu_Z(\mathbf{s}) = \mathbb{E}[Z(\mathbf{s})] = \sum_{z_i \in \Omega} z_i f_{Z(\mathbf{s})}(z_i), \tag{2.41}$$

where $f_{Z(\mathbf{s})}(\cdot)$ is the probability mass function (pmf) for $Z(\mathbf{s})$.

The covariance function or second moment of stochastic process $\{Z(\mathbf{s})\}$ is defined as

$$\begin{aligned} C_Z(\mathbf{s}, \mathbf{t}) &= \mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{t})], \\ &= \mathbb{E}\Big[(Z(\mathbf{s}) - \mu_Z(\mathbf{s}))(Z(\mathbf{t}) - \mu_Z(\mathbf{t}))\Big], \end{aligned} \tag{2.42}$$

and measures the strength of the linear dependence between two random variables $Z(\mathbf{s})$ and $Z(\mathbf{t})$. The variance function is a special case of the covariance function where $\mathbf{s} = \mathbf{t}$ and is given by

$$\begin{aligned} \text{Var}[Z(\mathbf{s})] &= C_Z(\mathbf{s}, \mathbf{s}) \\ &= \text{Cov}[Z(\mathbf{s}), Z(\mathbf{s})] \\ &= \mathbb{E}[(Z(\mathbf{s}) - \mu_Z(\mathbf{s}))^2] \\ &= \sigma_Z^2(\mathbf{s}). \end{aligned} \tag{2.43}$$

The covariance function is symmetric since $C_Z(\mathbf{s}, \mathbf{t}) = C_Z(\mathbf{t}, \mathbf{s})$ and a non-negative definite function since it satisfies the condition $\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j C_Z(\mathbf{s}_i, \mathbf{s}_j) \geqslant 0$ for all positive integers, m; real-valued constants, $a_1, \ldots, a_m$; and spatial locations, $(\mathbf{s}_1, \ldots, \mathbf{s}_m)$.

In addition to the covariance function, the correlation function of the stochastic process $\{Z(\mathbf{s})\}$ is a scaled version of the covariance function given by

$$\rho_Z(\mathbf{s}, \mathbf{t}) = \text{Corr}[Z(\mathbf{s}), Z(\mathbf{t})] = \frac{C_Z(\mathbf{s}, \mathbf{t})}{\sqrt{C_Z(\mathbf{s}, \mathbf{s}) C_Z(\mathbf{t}, \mathbf{t})}}, \tag{2.44}$$

and measures the strength of the linear association between the two random variables $Z(\mathbf{s})$ and $Z(\mathbf{t})$ and takes values between -1 and 1, i.e., $-1 \leqslant \rho_Z(\mathbf{s}, \mathbf{t}) \leqslant 1$ for all pairs $\mathbf{s}, \mathbf{t} \in D$.

2.4.1.2 Stationarity of the geostatistical process

Stationarity of the geostatistical process $\{Z(\mathbf{s})\}$ occurs when it has the same characteristics at any location, such as a constant mean or variance. There are two types of stationarity conditions and these are defined below.

A geostatistical process $\{Z(\mathbf{s})\}$ is strictly stationary when the process can be moved in space and it stays the same, i.e.,

$$f(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)) =_d f(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_m + \mathbf{h})), \tag{2.45}$$

where $=_d$ means equal in distribution. Here, $\mathbf{h}$ is a displacement or lag vector, which dictates the shifts in space. This means that the geostatistical process, $\{Z(\mathbf{s})\}$, has the same distribution for all locations within the spatial domain, thus we have constant mean $\mu_Z(\mathbf{s}) = \mu_Z$ and constant variance $\sigma_Z^2(\mathbf{s}) = \sigma_Z^2$; but does not mean the random variables $(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m))$ are independent. Furthermore, the bivariate distribution does not depend on the spatial location, i.e., $f(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{h})) =_d f(Z(\mathbf{0}), Z(\mathbf{h}))$ for all $\mathbf{s}$ and $\mathbf{h}$. This means that the covariance function between any two points only depends on the distance and direction between them, and not on the actual locations themselves, i.e., $\text{Cov}[Z(\mathbf{s}), Z(\mathbf{s}, \mathbf{h})] = C_Z(\mathbf{s}, \mathbf{s} + \mathbf{h}) = C_Z(\mathbf{h})$ - it does not change over

space.

Strictly stationary is a rather restrictive condition so a geostatistical process can also be weakly stationary if

1. The mean is constant across space, i.e., $\mathbb{E}[Z(\mathbf{s})] = \mu_Z(\mathbf{s}) = \mu_Z$.
2. The covariance (and correlation) only depends on the distance and direction and not on the location as described above.

Weakly stationary is typically assumed in a spatial analysis; however, it assumes a constant mean in space, which will not be true for the majority of data sets. Therefore, a regression model is fitted to the data to account for a non-constant trend in the mean before assuming the spatial autocorrelation remaining in the data is stationary.

A further restriction is enforced on the geostatistical process for simplification and is known as isotropy. A geostatistical process is weakly stationary if the covariance or correlation function only depends on the distance and direction, and isotropy further simplifies this condition to the covariance or correlation function only depending on the distance and not the direction or actual location. Mathematically this is defined as

$$C_Z(h) = C_Z(||\mathbf{h}||), \tag{2.46}$$

where $h = ||\mathbf{h}||$ is the Euclidean distance between any two elements of $\mathbf{h}$. In addition, since the covariance is isotropic, the correlation function is also isotropic since it is just a scaled version of the covariance.

2.4.1.3 Gaussian processes and covariance functions

The underlying spatial structure of the data can be specified in many ways, but one of the most common types of geostatistical processes is the Gaussian process. The Gaussian process is completely specified by its first two moments (mean, variance/covariance/correlation) and a geostatistical process is a Gaussian process if the joint distribution of the random variables $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}$ have a multivariate normal (Gaussian) distribution.

The joint probability density function of the random variables $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m))^\top$ at m locations is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-\frac{m}{2}} \det(\mathbf{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_Z)^\top \mathbf{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}_Z) \right), \tag{2.47}$$

where $\boldsymbol{\mu}_Z = (\mu_Z(\mathbf{s}_1), \ldots, \mu_Z(\mathbf{s}_m))^\top$ and the (j,k)th element of the covariance matrix $\boldsymbol{\Sigma}$ is $\Sigma_{jk} = C_Z(||\mathbf{s}_j - \mathbf{s}_k||)$. Furthermore, if the Gaussian process is weakly stationary then the process is also strictly stationary since the Gaussian distribution is completely specified by its first two moments.

Since the covariance function is used to model the correlation between observations, numerous models have been proposed that utilise the covariance function in geostatistical processes. The most commonly-used model for its simplicity is the exponential covariance function, defined as

$$C_Z(h) = \begin{cases} \sigma^2 \exp(-h/\rho), & h > 0; \\ \tau^2 + \sigma^2, & h = 0, \end{cases} \tag{2.48}$$

where $h = ||\mathbf{h}||$, σ^2 is the variance, ρ is the spatial decay parameter that measures how quickly the covariance decays to zero and τ^2 quantifies the amount of non-spatial variation or measurement error in the data.

The Gaussian covariance function is another covariance function, which is smoother compared to the exponential, but slightly more complicated. It is defined as

$$C_Z(h) = \begin{cases} \sigma^2 \exp(-(h/\rho)^2), & h > 0; \\ \tau^2 + \sigma^2, & h = 0. \end{cases} \tag{2.49}$$

There are many other covariance functions, including the power exponential, spherical exponential and wave exponential covariance functions (Diggle & Ribeiro, 2007). These are not shown as they are not considered in this thesis. Instead, the exponential covariance function is utilised due to its simplicity.

2.4.1.4 Maximum likelihood estimation

In classical parameter estimation, the geostatistical process is modelled as

$$Z(\mathbf{s}) = \mu_Z(\mathbf{s}) + \epsilon_Z(\mathbf{s}), \tag{2.50}$$

which is the trend plus error. The mean $\mu_Z(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta}$ is a linear combination of p covariates, that is, $\mathbf{x}(\mathbf{s})^\top = (\mathbf{x}_0, \mathbf{x}_1(\mathbf{s}), \ldots, \mathbf{x}_p(\mathbf{s}))$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. Since data are measured at m locations, the mean for geostatistical data $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))^\top$ is defined as

$$\boldsymbol{\mu}_Z = (\mu_Z(\mathbf{s}_1), \ldots, \mu_Z(\mathbf{s}_m)) = \mathbf{X}\boldsymbol{\beta}, \tag{2.51}$$

where $\mathbf{X}$ is the $m \times p$ design matrix of covariates for all m locations. The errors for all

m locations $\boldsymbol{\epsilon}_Z = (\epsilon_Z(\mathbf{s}_1), \ldots, \epsilon_Z(\mathbf{s}_m))^\top$ are modelled as

$$\boldsymbol{\epsilon}_Z \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.52}$$

where the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a function of the covariance parameters $\boldsymbol{\theta} = (\sigma^2, \tau^2, \rho)$, and its ijth element is the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$ as given by $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = C_Z(||\mathbf{s}_i - \mathbf{s}_j||)$, which is a covariance function that is weakly stationary and isotropic, such as the exponential covariance function.

Therefore, for geostatistical data $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))^\top$ the Gaussian geostatistical model considered is

$$\mathbf{z} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2.53}$$

where the mean function is a linear combination of known covariates $\mathbf{X}$ with associated regression parameters $\boldsymbol{\beta}$, and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as described above. For the exponential covariance function, it can be written in matrix form as

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2 \exp(-\mathbf{D}/\rho) + \tau^2 \mathbf{I}, \tag{2.54}$$

where $\mathbf{D}$ is an Euclidean distance matrix with ijth elements described by $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ and $\mathbf{I}$ is a $m \times m$ identity matrix. As this is a distance matrix, the diagonal elements are zero corresponding to $h = d_{ii} = 0$ and have values $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = \sigma^2 + \tau^2$, and non-diagonal elements have $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = \sigma^2 \exp(-d_{ij}/\rho)$. Furthermore, this matrix is positive definite and hence invertible.

The parameters $(\boldsymbol{\beta}, \sigma^2, \tau^2, \rho)$ are estimated by maximum likelihood by maximising the log-likelihood function of $\mathbf{z}$ based on the multivariate Gaussian distribution given in equation (2.47). Firstly, the log-likelihood function is differentiated with respect to each of the parameters and setting each derivative equal to zero. Secondly, the second derivative is calculated to ensure the estimator is indeed a maximum. The log-likelihood function is therefore given by (with unnecessary constants removed)

$$\begin{aligned} \ln(f(\mathbf{z})) \quad \propto \quad & -\frac{1}{2} \ln[\det(\sigma^2 \exp(-\mathbf{D}/\rho) + \tau^2 \mathbf{I})] \\ & -\frac{1}{2} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^\top (\sigma^2 \exp(-\mathbf{D}/\rho) + \tau^2 \mathbf{I})^{-1} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{2.55}$$

In order to aid estimation, the transformation $\nu^2 = \tau^2/\sigma^2$ can be applied, where the parameter ν^2 is known as the noise-to-signal ratio. Applying this transformation changes the log-likelihood function to

$$\begin{aligned}\ln(f(\mathbf{z})) \quad \propto \quad & -\frac{1}{2}m\ln(\sigma^2) - \frac{1}{2}\ln[\det(\exp(-\mathbf{D}/\rho) + \nu^2\mathbf{I})] \\ & -\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^\top(\exp(-\mathbf{D}/\rho) + \nu^2\mathbf{I})^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{2.56}$$

The log-likelihood function can be further simplified by rewriting the covariance function $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ from $\sigma^2\exp(-\mathbf{D}/\rho) + \tau^2\mathbf{I}$ to $\sigma^2\mathbf{V}(\rho,\nu^2)$ where $\mathbf{V}(\rho,\nu^2) = \exp(-\mathbf{D}/\rho) + \nu^2\mathbf{I}$. Therefore, to obtain the estimates for parameters $(\boldsymbol{\beta},\sigma^2)$ the log-likelihood function in equation (2.56) is differentiated with respect to these parameters, set to equal zero and solved to give the following.

For $\boldsymbol{\beta}$, only the parts of the log-likelihood containing $\boldsymbol{\beta}$ are utilised to give

$$\begin{aligned}\ln(f(\mathbf{z})) &= \frac{2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{z}}{2\sigma^2} - \frac{\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{X}\boldsymbol{\beta}}{2\sigma^2} \\ \frac{d\ln(f(\mathbf{z}))}{d\boldsymbol{\beta}} &= \frac{\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{z}}{\sigma^2} - \frac{\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{X}\boldsymbol{\beta}}{\sigma^2}.\end{aligned} \tag{2.57}$$

Setting this equal to 0 and solving for $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}}(\rho,\nu^2) = (\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}(\rho,\nu^2)^{-1}\mathbf{z}. \tag{2.58}$$

For σ^2, only the parts of the log-likelihood function containing σ^2 are utilised to give

$$\begin{aligned}\ln(f(\mathbf{z})) &= \frac{-m\ln(\sigma^2)}{2} - \frac{(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}(\rho,\nu^2)^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ \frac{d\ln(f(\mathbf{z}))}{d\sigma^2} &= \frac{-m}{2\sigma^2} + \frac{(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}(\rho,\nu^2)^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})}{2(\sigma^2)^2}.\end{aligned} \tag{2.59}$$

Setting this equal to 0 and solving for σ^2 gives

$$\hat{\sigma^2}(\boldsymbol{\beta},\rho,\nu^2) = \frac{1}{m}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}(\rho,\nu^2)^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}). \tag{2.60}$$

As the maximum likelihood estimator is biased as with standard linear model theory, the alternative is considered instead

$$\hat{\sigma^2}(\boldsymbol{\beta},\rho,\nu^2) = \frac{1}{m-p}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}(\rho,\nu^2)^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}), \tag{2.61}$$

where p is the number of parameters in the mean model.

For ρ and ν^2, differentiation of the log-likelihood does not work as both these parameters are contained within the inverted covariance matrix $\mathbf{V}(\rho, \nu^2)^{-1}$, so no closed form exists. Therefore, the estimates for $\boldsymbol{\beta}$ and σ^2 ($\hat{\boldsymbol{\beta}}(\rho, \nu^2), \hat{\sigma^2}(\hat{\boldsymbol{\beta}}, \rho, \nu^2)$) are plugged into the log-likelihood function to obtain a reduced form known as the profile/reduced likelihood that has to be maximised via numerical methods, such as grid searching. The profile likelihood for (ρ, ν^2) is given by

$$\ln(f(\mathbf{z})) \propto \frac{m}{2} \ln(\hat{\sigma^2}(\hat{\boldsymbol{\beta}}, \rho, \nu^2)) - \frac{1}{2} \ln[\det(\mathbf{V}(\rho, \nu^2))], \tag{2.62}$$

and the final maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}}(\hat{\rho}, \hat{\nu^2}) = (\mathbf{X}^\top \mathbf{V}(\hat{\rho}, \hat{\nu^2})^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}(\hat{\rho}, \hat{\nu^2})^{-1} \mathbf{z} \tag{2.63}$$

$$\hat{\sigma^2}(\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\nu^2}) = \frac{1}{m-p} (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\rho}, \hat{\nu^2}))^\top \mathbf{V}(\hat{\rho}, \hat{\nu^2})^{-1} (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\rho}, \hat{\nu^2})). \tag{2.64}$$

2.4.1.5 Spatial prediction

The objective of a geostatistical analysis is to not only identify and understand the spatial pattern in the data, but to also predict the process at unmeasured location $\mathbf{s}_0$. An approach called kriging was proposed by D. G. Krige, who worked in the South African mining industry in 1955 (Krige, 1951). Kriging is based on obtaining the best linear unbiased prediction (BLUP) for the process at new locations $\{Z(\mathbf{s}_0)\}$, given the observed data $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))^\top$. The BLUP can be obtained by choosing values for $(a_0, \mathbf{a})$ that minimise the mean squared prediction error (MSPE) defined by

$$\text{MSPE} = \mathbb{E}[(Z(\mathbf{s}_0) - P_{\mathbf{z}}(\mathbf{s}_0))^2], \tag{2.65}$$

where $P_{\mathbf{z}}(\mathbf{s}_0) = a_0 + \sum_{j=1}^{m} a_j Z(\mathbf{s}_j)$ is the linear prediction operator, a_0 is some constant, and $\mathbf{a} = (a_1, \ldots, a_m)$ are prediction weights.

There are two types of kriging predictors: the ordinary kriging predictor and the universal kriging predictor. The ordinary kriging predictor assumes a constant mean; whereas the universal kriging predictor allows for a non-constant mean which is much more realistic.

For the universal kriging predictor, assume there is a non-constant mean for the data such as $\mathbb{E}[\mathbf{Z}] = \mathbf{X}\boldsymbol{\beta}$, so the observed data are distributed as $\mathbf{z} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Then the combination of the unobserved and observed data have the following distribution

$$\mathbf{Z}^* = \begin{pmatrix} Z(\mathbf{s}_0) \\ \mathbf{Z} \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mathbf{x}_0\boldsymbol{\beta} \\ \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} C_Z(\mathbf{0}, \boldsymbol{\theta}) & \boldsymbol{c}_Z(\mathbf{s}_0, \boldsymbol{\theta})^\top \\ \boldsymbol{c}_Z(\mathbf{s}_0, \boldsymbol{\theta}) & \boldsymbol{\Sigma}(\boldsymbol{\theta}) \end{pmatrix} \right), \tag{2.66}$$

where $\mathbf{x}_0$ is the vector of covariates at the unobserved location $\mathbf{s}_0$, and $\boldsymbol{c}_Z(\mathbf{s}_0, \boldsymbol{\theta})$ is the vector of covariances at the unobserved location $\mathbf{s}_0$. To obtain the universal kriging

predictor for the value at the unobserved location $\mathbf{s}_0$ given the observed data $\mathbf{z}$, we make use of the conditional distribution property of a multivariate Gaussian distribution where the conditional distribution of $Z(\mathbf{s}_0)|\mathbf{Z}$ is given by

$$Z(\mathbf{s}_0)|\mathbf{Z} \sim \mathrm{N}(\mathbb{E}[\widehat{Z(\mathbf{s}_0)}|\mathbf{Z}], \mathrm{Var}[\widehat{Z(\mathbf{s}_0)}|\mathbf{Z}]), \tag{2.67}$$

and

- $\mathbb{E}[\widehat{Z(\mathbf{s}_0)}|\mathbf{Z}] = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \boldsymbol{c}_Z(\mathbf{s}_0, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}})$,
- $\mathrm{Var}[\widehat{Z(\mathbf{s}_0)}|\mathbf{Z}] = C_Z(\mathbf{0}, \hat{\boldsymbol{\theta}}) - \boldsymbol{c}_Z(\mathbf{s}_0, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{c}_Z(\mathbf{s}_0, \hat{\boldsymbol{\theta}})$, which allows for any uncertainty in the prediction. The hat notation here $(\widehat{Z(\mathbf{s}_0)})$ denotes that these are estimates of the mean and variance at unmeasured locations $\mathbf{s}_0$.

The above formula can be extended to predict the process at N unmeasured prediction locations $\mathbf{s}^* = (\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*)$ for random variables $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_N^*))$ as

$$\mathbf{Z}^*|\mathbf{Z} \sim \mathrm{N}(\mathbb{E}[\widehat{\mathbf{Z}^*}|\mathbf{Z}], \mathrm{Var}[\widehat{\mathbf{Z}^*}|\mathbf{Z}]), \tag{2.68}$$

where

$$\begin{aligned} \mathbb{E}[\widehat{\mathbf{Z}^*}|\mathbf{Z}] &= \mathbf{X}^*\hat{\boldsymbol{\beta}} + \boldsymbol{c}_Z(\mathbf{s}^*, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{Z} - \mathbf{X}^*\hat{\boldsymbol{\beta}}), & (2.69)\\ \mathrm{Var}[\widehat{\mathbf{Z}^*}|\mathbf{Z}] &= \boldsymbol{\Sigma}^*(\hat{\boldsymbol{\theta}}) - \boldsymbol{c}_Z(\mathbf{s}^*, \hat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{c}_Z(\mathbf{s}^*, \hat{\boldsymbol{\theta}}), & (2.70) \end{aligned}$$

where $\boldsymbol{\Sigma}^*(\hat{\boldsymbol{\theta}})$ is a $N \times N$ variance matrix for the N prediction locations, $\mathbf{X}^*$ is a matrix of covariates at the N prediction locations, and $\boldsymbol{c}_Z(\mathbf{s}^*, \hat{\boldsymbol{\theta}})$ is an $N \times m$ covariance matrix between the prediction and observed locations. This type of plug in prediction can result in 95% prediction intervals that are too narrow due to the uncertainty in $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ not being taken into account, meaning that the confidence intervals contain the true value less than 95% of the time. This is rather troublesome as it means that the confidence intervals are not taking the inherent uncertainty in the predictions into account.

2.4.1.6 Bayesian methods for geostatistical processes

The difference between classical parameter estimation, i.e., maximum likelihood, and Bayesian methods is that Bayesian methods do not only provide a point estimate, but also provide information about the entire distribution of the parameters given the observed data. This distribution is known as the posterior distribution as given by Bayes Theorem in equation (2.26) in Section 2.3. A point estimate, such as the mean or median, of the posterior distribution can be taken and a 95% credible interval (or uncertainty interval) can be obtained by calculating the $[2.5, 97.5]$ percentiles of the

posterior. As discussed in Section 2.3, the most common approach is to assume a non-informative or vague prior that contains little information about the parameters. Working within a Bayesian setting performs prediction correctly by allowing for the variation in $(\boldsymbol{\beta}, \boldsymbol{\theta})$ when doing the prediction, rather than the plug-in approach utilised by maximum likelihood. However, Bayesian methods are much slower compared to maximum likelihood estimation due to the need to simulate entire distributions rather than just obtaining point estimates.

A general modelling framework is given by

$$\begin{aligned} Z(\mathbf{s}_i) &\sim \mathrm{N}(\mu_i, \tau^2) \quad \text{for } i = 1, \ldots, m, \\ \mu_i &= \mathbf{x}_i^\top \boldsymbol{\beta} + \phi(\mathbf{s}_i), \\ \boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_m)) &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \end{aligned} \tag{2.71}$$

where $\boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_m))$ for all locations are known as random effects and allow for any unmeasured spatial autocorrelation in the data after the covariate effects have been accounted for. This is essentially a generalised linear mixed model, where the random effects are spatially correlated and are modelled by a Gaussian geostatistical process, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is defined through an isotropic covariance function, such as the exponential correlation function. McMC is then used to sample values for each of the parameters in order to provide the entire distribution, which is discussed in Section 2.3.

When predicting the geostatistical process at unmeasured locations, the random effects also have to be generated at the N prediction locations given as $\boldsymbol{\phi}^* = (\phi(\mathbf{s}_1^*), \ldots, \phi(\mathbf{s}_N^*))$. This is done using the same multivariate Gaussian theory as above, namely

$$\boldsymbol{\phi}^* \sim \mathrm{N}(\mathbb{E}[\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}], \mathrm{Var}[\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}]), \tag{2.72}$$

where the mean and variance are given by

$$\begin{aligned} \mathbb{E}[\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}] &= \boldsymbol{c}_Z(\mathbf{s}^*, \boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{\phi}, \\ \mathrm{Var}[\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}] &= \boldsymbol{\Sigma}^*(\boldsymbol{\theta}) - \boldsymbol{c}_Z(\mathbf{s}^*, \boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{c}_Z(\mathbf{s}^*, \boldsymbol{\theta}) \end{aligned} \tag{2.73}$$

coming from multivariate Gaussian theory as shown in equations (2.69) and (2.70).

2.4.2 Areal unit statistics

The second type of spatial data used in this thesis are areal unit data, where the study region of interest, D, is partitioned into a finite number of non-overlapping subregions $\{\mathcal{A}_i : i = 1, \ldots, m\}$, and data are observed at the subregion level. The subregions have the following characteristics

$$\begin{aligned} \cup_{i=1}^{m} \mathcal{A}_i &= D \quad \text{and} \\ \mathcal{A}_i \cap \mathcal{A}_j &= \emptyset \quad \text{for each } i \neq j, \end{aligned} \tag{2.74}$$

thus an areal process is a stochastic process defined by

$$\{Z_i = Z(\mathcal{A}_i) : i = 1, \ldots, m\}, \tag{2.75}$$

where Z_i is a random variable representing the quantity of interest at subregion (areal unit) $\mathcal{A}_i$.

As discussed above, the main goal in areal unit analysis is to understand the spatial pattern in the data by producing maps and also to estimate the effects of a predictor on a response, while taking into account any spatial dependence or spatial autocorrelation. Spatial dependence is characterised by a proximity matrix, $\mathbf{W}$, which defines how the subregions are related to each other, be it in terms of the distance between the regions or in terms of adjacency, i.e., which subregions neighbour each other. Let $\mathbf{W}$ denote a $m \times m$ matrix with w_{ij} denoting the proximity between subregion $\mathcal{A}_i$ and subregion $\mathcal{A}_j$. The diagonal elements of the matrix are zero as the distance between a subregion and itself is zero; a subregion cannot border itself. There are many approaches to specifying the nature of $\mathbf{W}$ and one approach is to utilise a binary specification, which assigns the value 1 to w_{ij} if subregions $(\mathcal{A}_i, \mathcal{A}_j)$ neighbour each other, and 0 otherwise.

Typically, neighbours are defined as such when two subregions share a common border (denoted by $i \sim j$). These neighbouring areal units are modelled as correlated, while the non-neighbours are conditionally independent given the remaining neighbours. Neighbours can also be defined in terms of the distance between the centre of the subregions and if subregion $\mathcal{A}_i$ is one of the k closest in terms of distance to subregion $\mathcal{A}_j$. Distance measures of proximity may be inappropriate for irregularly-shaped areal units, such as electoral wards, since they are not consistent in shape or size across the study region. In this case, the share-a-common-border approach is utilised.

$\mathbf{W}$ is used to define the spatial dependence between the m subregions; however, it does not indicate the strength of the spatial dependence present in the data. Moran's I statistic (Moran, 1950) is a measure of the linear association in areal unit data weighted

by the proximity between subregion $\mathcal{A}_i$ and subregion $\mathcal{A}_j$. It is an extension of Pearson's correlation coefficient and is defined as

$$I = \frac{m \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{w_{\bullet\bullet} \sum_{i=1}^{m}(Z_i - \bar{Z})^2}, \tag{2.76}$$

where $\bar{Z} = \sum_{i=1}^{m} Z_i/m$ is the average of the spatial process over all areal units, and $w_{\bullet\bullet}$ is the sum of all w_{ij}'s, i.e., $w_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}$. Moran's I statistic can take a range of positive and negative values between -1 and 1 and describes whether there is negative spatial autocorrelation, no spatial autocorrelation, or positive spatial autocorrelation. Negative spatial autocorrelation takes values between -1 and < 0 where a statistic of -1 indicates perfect dispersion between the areal units, i.e., dissimilar areal units are located next to each other. Positive spatial autocorrelation takes values between > 0 and 1 where a statistic of 1 indicates perfect correlation, i.e., similar areal units are clustered next to each other. No spatial autocorrelation occurs with a statistic of 0 and indicates the similarity and dissimilarity of areal units are randomly arranged.

The significance of spatial autocorrelation can be quantified using a permutation test, which is a non-parametric approach to testing the significance of a statistic. It provides a simple way of calculating the sampling distribution of a test statistic under the null hypothesis by calculating K different random permutations of the dataset. The hypotheses for Moran's I permutation test are

$$\begin{aligned} H_0 &- \text{no spatial association} \\ H_1 &- \text{some spatial association,} \end{aligned} \tag{2.77}$$

where some spatial association could relate to either positive or negative spatial autocorrelation. The observed Moran's I statistic , I_{obs}, is calculated, on the raw data. Moran's I statistics are then calculated on the K different random permutations of the data set, given by $I_1, \ldots, I_K$, then the estimated two-sided p-value for the test is given by

$$\frac{2}{K+1} \sum_{k=1}^{K} I(I_k > |I_{obs}|). \tag{2.78}$$

There are numerous ways of measuring spatial associations, such as Local Indicators of Spatial Association (LISA) and Geary's contiguity ratio (Bivand et al., 2013), but these are not considered in this thesis as Moran's I is the most common approach and is widely used in the analysis of the geographic differences in health outcomes.

2.4.2.1 Constructing spatial dependence

Gaussian Markov Random Fields (GMRFs) are multivariate Gaussian distribution models that are used to construct dependence among the random variables of the areal process given in equation (2.75). The class of models used in this thesis are Conditional Autoregressive (CAR) models, which are the most common approach to modelling spatial autocorrelation in areal unit data. These CAR models are an extension of the autoregressive models seen in the branch of time series statistics that model short-term autocorrelation in temporal data after the trend and seasonal pattern has been removed.

CAR models are specified by a set of m univariate full conditional distributions given by

$$Z_i|\mathbf{Z}_{-i} \sim \mathrm{N}\left(\sum_{j=1}^{m} b_{ij} Z_j, \tau_i^2\right), \tag{2.79}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_m)^\top$, the mean function is a linear combination of the remaining random variables, b_{ij} are the regression coefficients, with $b_{ii} = 0$ for all i, and $\mathbf{Z}_{-i}$ denotes the set of random variables not containing the ith term. The joint distribution corresponding to these full conditionals is

$$f(Z_1, \ldots, Z_m) \propto \exp\left(-\frac{1}{2}\mathbf{Z}^\top \mathbf{K}^{-1}(\mathbf{I} - \mathbf{B})\mathbf{Z}\right), \tag{2.80}$$

where $\mathbf{B} = (b_{ij})$, $\mathbf{K} = \mathrm{diag}(\tau_1^2, \ldots, \tau_m^2)$ contains the variances, and $\mathbf{I}$ is the identity matrix of appropriate order. This joint distribution function for $\mathbf{Z}$ can also be written as

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{-1}), \tag{2.81}$$

where the inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1} = \mathbf{K}^{-1}(\mathbf{I} - \mathbf{B})$. This is a multivariate Gaussian distribution with mean zero and precision matrix $\mathbf{Q} = \mathbf{K}^{-1}(\mathbf{I} - \mathbf{B})$, and hence variance matrix $\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{K}$. The covariance matrix $\boldsymbol{\Sigma}$ is valid when the precision matrix $\mathbf{Q}$ is symmetric, otherwise $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}$ is not a valid variance matrix. The precision matrix is symmetric when

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \quad \forall\, i, j. \tag{2.82}$$

The most common specification is to set

$$b_{ij} = \frac{w_{ij}}{\sum_{i=1}^{m} w_{ij}}, \quad \tau_i^2 = \frac{\tau^2}{\sum_{i=1}^{m} w_{ij}} \tag{2.83}$$

as the neighbourhood matrix $\mathbf{W}$ is symmetric.

The intrinsic model (ICAR) is the simplest conditional autoregressive model proposed by Besag et al. (1991) and its univariate Gaussian conditional form is given by

$$Z_i|\mathbf{Z}_{-i} \sim \mathrm{N}\left(\frac{\sum_{j=1}^{m} w_{ij} Z_j}{\sum_{j=1}^{m} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{m} w_{ij}}\right). \tag{2.84}$$

The conditional expectation of Z_i is the mean of the random effects in neighbouring areas as $w_{ij} = 0$ for non-neighbouring areas, therefore it is only conditioning on areas that actually border. This means that neighbouring areas where $w_{ij} \neq 0$ are correlated and non-neighbouring areas are conditionally independent given the rest of $\mathbf{Z}$, i.e., $Z_i|\mathbf{Z}_{-i}$. The conditional variance is inversely proportional to the sum of the number of neighbours, i.e., the total number of neighbours. If strong spatial autocorrelation is present, the conditional variance uses the fact that areas with more neighbours have more information and thus can infer the value of the random effect. The variance parameter, τ^2, controls the amount of variation between the random effects and the conditional variance decreases with increasing number of neighbours. This is the simplest CAR specification as it does not take into account the strength of correlation in the data making this CAR model quite restrictive. Furthermore, it only models strong spatial correlation and is therefore not appropriate for weakly correlated data.

2.4.2.2 General modelling framework

These types of CAR models are typically implemented within a Bayesian framework, with inference based on Markov chain Monte Carlo (McMC) simulation. A general Bayesian hierarchical modelling framework for areal unit data is given by

$$\begin{aligned}
Z_i &\sim p(Z_i \mid \mu_i), \quad \text{for } i = 1, \ldots, m \\
G(\mu_i) &= \mathbf{z}_i^\top \boldsymbol{\beta} + \phi_i, \\
\phi_i|\boldsymbol{\phi}_{-i} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^{m} w_{ij}\phi_j}{\sum_{j=} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{m} w_{ij}}\right), \\
\boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta), \\
\tau^2 &\sim \text{Inverse-Gamma}(a, b),
\end{aligned} \tag{2.85}$$

where $p(Z_i \mid \mu_i)$ is a likelihood such as Poisson, $\mu_i = \mathbb{E}[Z_i]$ and $G(.)$ is a link function. Spatial autocorrelation comes in at the linear predictor level comprising a set of known covariates $\mathbf{z}_i$ with associated regression parameters $\boldsymbol{\beta}$ and a spatial random effect $\boldsymbol{\phi}$. The spatial random effects for all m areas are collectively denoted by $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$, and allow for any unmeasured spatial autocorrelation in the data after the covariate effects have been accounted for. This is essentially a generalised linear mixed model where the random effects are spatially autocorrelated and are mod-

elled by a CAR model. This Bayesian framework assumes the data are conditionally independent given the random effects, thus the data likelihood is given by

$$p(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\phi}, \tau^2) = \prod_{i=1}^{m} p(Z_i|\boldsymbol{\beta}, \phi_i), \tag{2.86}$$

and the joint prior distribution is given by

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\phi}, \tau^2) &= p(\boldsymbol{\beta})p(\boldsymbol{\phi}|\tau^2)p(\tau^2), \\ &= \mathrm{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \mathbf{V}_\beta)\mathrm{ICAR}(\boldsymbol{\phi}|\tau^2, \mathbf{W})\mathrm{IG}(\tau^2|a, b). \end{aligned} \tag{2.87}$$

The parameters of the prior distribution $(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta, a, b)$ are hyperparameters and are typically chosen to be vague and non-informative.

In areal unit data, the main aim is to model the spatial pattern in the mean and is typically performed in an ecological regression type setting. Typically, Poisson log-linear models are used to estimate the effect of a predictor on a response, for example estimating the effects of air pollution on heath via counts of mortality or morbidity. The modelling framework for the Poisson log-linear model is given by

$$\begin{aligned} Z_i &\sim \text{Poisson}(E_i R_i), \quad \text{for } i = 1, \ldots, m, \\ \ln(R_i) &= \mathbf{z}_i^\top \boldsymbol{\beta} + \phi_i, \\ \phi_i|\boldsymbol{\phi}_{-i} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^{m} w_{ij}\phi_j}{\sum_{j=}^{m} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{m} w_{ij}}\right), \\ \boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta), \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b), \end{aligned} \tag{2.88}$$

where the link function is a log-link function, E_i are the expected number of cases and form the offset for the model, and R_i is the risk of disease in areal unit i. The mean function is re-parameterised from $\mathbb{E}[Z_i] = \mu_i = E_i R_i$. As these models are implemented within a Bayesian setting, inference is based on McMC simulation using a combination of Gibbs sampling and Metropolis-Hastings, as described in Section 2.3.2.

2.4.2.3 Additional CAR models

The ICAR prior given by equation (2.84) only models strong spatial autocorrelation and can enforce too much spatial smoothness on the random effects. Therefore, numerous extensions have been proposed to overcome these issues.

The convolution model, also known as the BYM model, is an extension of the intrinsic model as it includes a second set of independent and identically distributed

random effects with mean zero and constant variance. The convolution specification is given by

$$\begin{aligned} \phi_i &= \phi_i^{(1)} + \phi_i^{(2)}, && (2.89) \\ \phi_i^{(1)}|\boldsymbol{\phi}_{-i}^{(1)} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^m w_{ij}\phi_j^{(1)}}{\sum_{j=1}^m w_{ij}}, \frac{\tau_1^2}{\sum_{j=1}^m w_{ij}}\right), && (2.90) \\ \phi_i^{(2)} &\sim \mathrm{N}(0, \tau_2^2), \end{aligned}$$

where the ICAR prior is represented by $\boldsymbol{\phi}_i^{(1)}$ and $\boldsymbol{\phi}^{(2)}$ represents independence between areas. The strength of spatial correlation is determined by the sizes of $(\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)})$ and overcomes the problem of the ICAR prior by letting the data choose the amount of spatial correlation in the data. The level of smoothness between the random effects is determined by the ratio of the two variances τ_1^2/τ_2^2. However, this model requires the fitting of two random effects for each area and only their sum is identifiable and reliably estimated.

Again, the convolution prior does not contain a parameter that specifically controls the level of spatial autocorrelation. The CAR model proposed by Cressie (1993) contains only one set of random effects, but an additional parameter ρ is specified to control the level of spatial autocorrelation. This CAR specification is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\rho\frac{\sum_{j=1}^m w_{ij}\phi_j}{\sum_{j=1}^m w_{ij}}, \frac{\tau^2}{\sum_{j=1}^m w_{ij}}\right). \quad (2.91)$$

This model is similar to the ICAR model, except the conditional mean is weighted by the level of spatial correlation in the data, with $\rho = 0$ corresponding to independence and $\rho = 1$ corresponds to strong spatial correlation, i.e., the ICAR model. One issue with this model is that, when ρ is zero there is no need for the conditional variance to be inversely proportional to the number of neighbours since the areas are independent.

Therefore, another CAR model was proposed by Leroux et al. (1999) that allows for varying degrees of spatial autocorrelation in the data. The Leroux model has the form

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\rho\sum_{j=1}^m w_{ij}\phi_j}{\rho\sum_{j=1}^m w_{ij} + (1-\rho)}, \frac{\tau^2}{\sum_{j=1}^m w_{ij} + (1-\rho)}\right). \quad (2.92)$$

Again, ρ controls the level of spatial correlation in the data, with $\rho = 0$ corresponding to spatial independence with mean zero and constant variance and $\rho = 1$ corresponding to strong spatial correlation (ICAR model). This is a similar model to the Cressie model, but it has the added flexibility that the conditional variance is no longer directly divided by the total number of neighbours. This model captures global spatial autocorrelation and is the CAR model used in this thesis due to its modelling flexibility.

2.5 Standardisation

Populations are inherently heterogeneous in terms of their sociodemographic structure (e.g., age, gender, education, ethnicity), and in terms of numerous personal and environmental factors related to health, such as diet and access to amenities. A population can therefore be viewed as a collection of different subgroups. Any overall measurements and statistics describing the population are often known as crude measurements. Crude measures are averages of particular subgroups weighted by the total size of the subgroups and the larger the subgroup, the more it will influence the crude measure.

Let N denote the size of the population consisting of a specific number of age groups, or strata. Each stratum contains a proportion of the total population, n_i, where i relates to the total number of strata. Within a specific time frame, each stratum will experience a certain number of deaths, d_i. The total size of the population is therefore Σn_i, the total number of deaths, D, is Σd_i, and the crude death or mortality rate is D/N. Crude rates are the simplest way of obtaining population summaries. However, the main disadvantage with crude rates is that they do not take into account the heterogeneity in the population. For example, if comparing the death rate in an older population to the death rate in a younger population, the older population would likely have more deaths and therefore a higher death rate, so the death rate is influenced by the age structure of the population.

Comparing rates across populations or over different time periods results in rates that are not directly comparable. This is due to the populations differing in composition so that what is observed may be attributed to these differences, such as the age structure in the previous example. Standardisation is a procedure which allows for the comparison of different populations or subgroups by taking into account the population or subgroup composition. There are two main standardisation methods: direct (or external) standardisation and indirect (or internal) standardisation. Indirect standardisation is the most common approach as it compares the actual number of events in a local population or area (e.g., Glasgow) with the expected number of events when strata-specific rates (e.g., based on the age and sex distribution) in a reference population (e.g., whole of Scotland) are applied to the local population or area. This produces a standardised mortality ratio (SMR), which can then be used to compare the local populations to the reference (or standard) population. However, SMRs cannot be directly compared to one another - only to the reference, and indirect standardisation can only be used if the strata-specific rates in the standard population are known.

Direct (or external) standardisation applies the local strata-specific rates to the standard population. This allows for direct comparison between local populations, for example, comparing the incidence of cardio-respiratory disease across regions in Scot-

land, and allows for differing age and sex (or other demographic factors) structures in each of the local areas. In contrast to the indirect method, direct standardisation requires strata-specific rates for the local population and not the standard population. Furthermore, if the heath outcome being studied is relatively rare, the direct method becomes unstable due to few events occurring in the stratum of the local population. As a result, the indirect method tends to be more commonly-used, as it requires strata-specific rates for the reference population and not the local population.

2.5.1 Direct standardisation

By applying age and sex specific mortality rates of the population(s) under study to the age and sex distribution of the reference population, direct standardisation ensures that the mortality rate is independent of differences in the age and sex distribution between populations. Furthermore, directly age-sex standardised mortality rates are the rates these populations would have experienced if they had the same age-sex distribution as the reference population (Roalfe et al., 2008).

The directly standardised rate is calculated by dividing the total expected number of cases in the standard population by its population size

$$\text{Standardised rate} = \frac{\sum_{ij} N_{ij}\hat{p}_{ij}}{N}, \tag{2.93}$$

where $i = 1, \ldots, 19$ age groups, $j = 1, 2$ sexes, N_{ij} is the number of people in the standard population in age group i and sex j, $\sum_{ij} N_{ij} = N$ is the total population from the standard, p_{ij} is the age-sex rate in the study population, and $\hat{p}_{ij} = r_{ij}/n_{ij}$ is the age-sex specific crude mortality rate in the study population, with the number of deaths denoted by r_{ij} and the number of people in each age-sex group of the study population denoted by n_{ij}.

The most common reference population is the European standard population [1] as it gives more weight to older age groups, which best reflects an ageing population, especially for the UK. The European standard population is a hypothetical population used for comparing different countries across Europe comprising equal numbers of males and females within each age band and totalling to a population of 200,000. Table 2.1 displays the age-sex distribution of the European standard and shows relatively high proportions of people in older aged groups, with the majority of the population falling into the working-age group. There are other standard populations that can be used depending on the age-sex structure of the study population. For example, a standard population comprising a high proportion of young people would be suitable for making comparisons with African populations. In addition, there is a World standard

[1] Most recent 2012 version available from: www.isdscotland.org

population that is based on the populations of 46 countries for producing rates that can be compared across the globe. However, this World standard population is a younger population compared to the European standard, thus not appropriate for the age-sex distribution of the population in Scotland.

Table 2.1: *Age-sex distribution of the European standard population based on the 2013 version.*

Age group	Male	Female
0-4	5,000	5,000
5-9	5,500	5,500
10-14	5,500	5,500
15-19	5,500	5,500
20-24	6,000	6,000
25-29	6,000	6,000
30-34	6,500	6,500
35-39	7,000	7,000
40-44	7,000	7,000
45-49	7,000	7,000
50-54	7,000	7,000
55-59	6,500	6,500
60-64	6,000	6,000
65-69	5,500	5,500
70-74	5,000	5,000
75-79	4,000	4,000
80-84	2,500	2,500
85-89	1,500	1,500
90+	1,000	1,000
Total	100,000	100,000

2.5.2 Indirect standardisation

One of the issues with direct standardisation is that the rates become unstable if there are too few events in the age-sex groups. If there are many strata with zero rates, it results in rates that are susceptible to being heavily influenced by random variability, rendering direct standardisation unsatisfactory. Indirect standardisation avoids the issue of imprecise estimates by applying age-sex specific rates from the reference population to the age-sex structure of the study population. Indirect standardisation compares the observed number of deaths in the study population to the expected number of deaths, i.e., the number of deaths that would be expected if the study population bore the same age-sex structure of the reference population. The SMR is,

therefore, the ratio of the observed number of deaths to the expected number of deaths.

The reference population used to perform internal standardisation in this thesis was the population of West Central Scotland, i.e., the age-sex distribution of the study region over all subregions. The West Central Scotland population was chosen because the national Scottish mortality rates do not wholly reflect the mortality rates in West Central Scotland due to the fact that the Glasgow conurbation has the highest levels of deprivation in Scotland, thus it is a more representative reference population compared to Scotland as a whole.

Let Y_k denote the observed number of events (e.g., deaths) in subregion k and let E_k denote the expected number of events in subregion k. The expected number of events in subregion k is calculated as

$$E_k = \sum_{r=1}^{38} N_{kr}\gamma_r, \tag{2.94}$$

where N_{kr} denotes the number of people in age-sex group r in subregion k, and γ_r denotes the rate of events in the standard population in age-sex group r and is given by the number of events in age-sex group r divided by the population in age-sex group r.

Chapter 3

Review of air pollution and health studies

3.1 Introduction

Quantifying the impact of air pollution on ill health is conducted using three main types of studies, namely time series studies, cohort studies and areal unit studies. Ecological time series studies, such as Omori et al. (2003), and Moolgavkar et al. (2013), are the most common type of study design due to being quick and inexpensive to implement since the data are readily available. These studies examine the effects of short-term (acute) exposure on human health by regressing routinely available air pollution and disease data collected at daily intervals. A detailed systematic review and meta-analysis of the associations between short-term exposure to nitrogen dioxide and health can be found in Mills et al. (2015). The disease data comprise population level summaries (typically counts) of mortality (Kinney & Ozkaynak, 1991) or morbidity, such as hospital admissions (Willocks et al., 2012), for a number of common disease such as respiratory (Atkinson et al., 2001) and cardiovascular conditions (Larrieu et al., 2007).

The long-term (chronic) health impact of air pollution is most often estimated from cohort studies (Brunekreef, 2007; Cesaroni et al., 2014; Dockery et al., 1993; Jerrett et al., 2009; Pope III et al., 2002, 1995), which make use of individual-level air pollution and disease data. However, cohort studies are expensive and time consuming to implement due to the length of follow-up required for monitoring the health status of the cohort. This has led to spatial ecological study designs being used (Haining et al., 2010; Lee et al., 2009; Maheswaran et al., 2005a), which make use of routinely available small area data, such as from the Scottish Neighbourhood Statistics (`http://www.sns.gov.uk/`) database, and the Health and Social Care Information Centre (`http://www.hscic.gov.uk/`). Due to their ecological nature these studies cannot be used to determine individual-level causality, but they contribute to and independently corroborate the body of evidence provided by cohort studies.

Spatial ecological studies are the main focus of this thesis, and are based on partitioning the study region into m contiguous small areas determined by administrative boundaries, such as electoral wards or census tracts, with smaller areas considered to comprise more homogeneous populations compared to larger areas; for example, in terms of social characteristics. For each small area, the response is the number of disease cases in a fixed time period, such as the number of deaths due to respiratory disease in one year. These disease cases are adjusted for varying population demographics across the study region using indirect standardisation, and then regressed against air pollution concentrations and other confounders, such as socio-economic deprivation. Typically, Poisson log-linear models are used to estimate the pollution-health effect, and any residual spatial autocorrelation in the data is accounted for by introducing a set of spatially autocorrelated random effects into the model. This residual spatial autocorrelation could be due to numerous factors, including unmeasured confounding (where an important spatially correlated variable is not included in the model or is unknown), neighbourhood effects (where the behaviour of subjects is influenced by surrounding subjects), and grouping effects (where subjects of similar characteristics group together).

This literature review focuses on air pollution and health studies that have incorporated an ecological areal unit design, and the statistical challenges faced when developing a concrete model to assess the association between air pollution and health. These challenges include the difficulties of defining the study area and thus the size of the small areas with regards to the abundance of health data that are available, the risk of ecological bias, misestimation of exposure to air pollution and other covariates, and modelling residual spatial autocorrelation. These factors are important when developing a statistical model as they all affect the complexity of the models and thus there is often a trade-off between model simplicity and ensuring the model accurately reflects the nature of the processes at play.

This review provides background information on the methods utilised in Chapters 4 and 5, while discussing the aforementioned statistical challenges. Section 3.2 describes further the nature of ecological studies, their benefits and shortcomings; while Section 3.3 outlines the typical disease, air pollution and covariate data used in ecological areal unit studies, along with the standard modelling approach. Section 3.4 provides a brief overview of studies conducted in Scotland and other parts of the world, while providing an in depth critique of small area studies. The remainder of the chapter focuses on the limitations of areal unit studies, where Section 3.5 discusses the concept of ecological bias and studies that try to mitigate its effects. The estimation of air pollution exposure is discussed Section 3.6.

3.2 Ecological studies

Spatial ecological or small area studies seek to analyse the geographical pattern of disease in terms of demographic, environmental, deprivation, as well as other important factors. Spatial ecological studies aim to understand which spatially varying environmental factors influence the risk of disease (Elliott & Savitz, 2008), and in this thesis the spatially varying factor relates to air pollution.

Environmental exposures, such as air pollution, are largely influenced by location, but also by meteorological factors. Exposure at the individual level to air pollution is determined by geographical factors such as home, work, and school, but most importantly, how one moves from place to place through the air pollution surface and the time spent outside. Incidentally, this is determined by the individual in terms of age, sex, social class, job and therefore, level of income, as well as how an individual travels between places. For example, a child will spend most of their time at school while being active indoors and outdoors, an adult who works likely spends most of their time at their job, and a retired individual likely spends most of their time at home. While it is important to establish individual levels of exposure to air pollution, it is often impractical and not cost-effective as it would require individual-based air pollutant monitors for a large cohort of the at-risk population. Cohort studies therefore rely on proxy measures of exposure, which can either be simple, for example, by measuring the distance from a point source (Elliott et al., 1996), or distance to the nearest road (Hoek et al., 2002; Wilkinson et al., 1999), or they can be more complex, using, for example, dispersion modelling (Havard et al., 2009).

Spatial ecological studies differ from cohort studies in that they are carried out at the population level on aggregated data for both the outcome and exposure, rather than at the individual level. Furthermore, ecological studies, in general, are more convenient and are not expensive to conduct due to the availability and ease of access to population-level and routine data. This type of study can be cross-sectional when considering only a single time point, such as all disease cases within a year, or longitudinal when considering the change in disease and exposure over time. Spatial ecological studies can be descriptive (such as disease mapping; MacNab et al., 2006), which seeks to describe the distribution and spatial pattern of an outcome, such as cardio-respiratory mortality, across the chosen geographical area and study period. This type of descriptive study helps generate further research questions, which can then be studied in a more formal framework (Bailey et al., 2005). Therefore, analytical spatial ecological studies aim to investigate the relationship between an exposure and outcome, while taking into account any residual spatial autocorrelation.

One of the main issues with ecological studies is that they may not be able to

measure information on important risk factors (or confounders) thought to be associated with the disease under study, such as smoking levels when considering cardiorespiratory diseases. This is due to the fact that data in general are not collected for the purpose of the chosen study, but for other purposes such as surveillance. Likewise, data on the exposure and outcome are usually collected in different ways. For example, air pollution is measured at single points in space, whereas the outcome can be individual hospital records, and this can inadvertently bias the results. Furthermore, the way in which data are collected can differ systematically over time, for example, when air pollution monitors are upgraded, removed, or change location. While ecological studies are a powerful tool for investigating quickly and efficiently the relationship between an exposure and outcome at the population level, they cannot be used to infer associations at the individual level, otherwise this is known as the ecological fallacy as is discussed further in Section 3.2.

3.3 Study design and data

The air pollution, disease and covariate data are typically recorded at the monthly or annual level for each administrative unit in the chosen study region. The size of the administration units varies from study to study, ranging from small census enumeration districts, comprising 400 inhabitants on average, to large local unitary authorities, comprising 200,000 inhabitants on average. It is clear that the smaller the area under study the more homogeneous the populations living within the areas are considered to be. The analyses presented in Chapters 5 and 6 are based on data from West Central Scotland, where the size of the administration units, known as data zones, comprise 800 inhabitants on average. The air pollution, disease and covariate data used in these studies are described below, firstly with a definition on the ways in which disease can be studied.

3.3.1 Frequency of disease

The occurrence of morbidity and mortality varies over time, across space and between different population groups, such as the working age population and the older age population. Therefore, it is important to be able to quantify the frequency of disease in order to allow the study of these events when seeking to develop an intervention, when seeking to prevent disease and promote health, or when seeking to identify a causal relationship between an exposure and an outcome, such as between air pollutants and health.

There are two main measures of disease frequency: prevalence and incidence. Prevalence is defined as the number of existing diseases cases within a defined population within a specified time period. Prevalence is an important measure as it allows the as-

sessment of the public health impact of a specific disease within a population, however it cannot be used to establish any causal relationships (Bailey et al., 2005), since it is studied at the population level. This is mainly due to not being able to fully establish the factors that lead to the disease and the factors that exacerbate these conditions. However, this can be mitigated by refining the study population, for example in terms of age since air pollution will have different effects at different ages. Incidence is a more useful measure as it only looks at new cases of disease. Incidence is defined as the number of new disease cases within a defined population and time period, see, for example, Atkinson et al. (2013) who analysed the incidence of cardiovascular disease over a 5-year period. Incidence is studied in Chapter 6.

3.3.2 Disease data

In spatial ecological studies, the mortality or morbidity data are available as aggregated counts for each of the m non-overlapping subregions $\{\mathcal{A} : i = 1, \ldots, m\}$ within the study region of interest, D. These disease data are denoted by $\mathbf{Y} = (Y_1, \ldots, Y_m)$, where Y_i represents the number of disease cases within areal unit i. Disease count data are available from Government departments, such as the National Records of Scotland (NRS, https://www.nrscotland.gov.uk/), whereby individual-level information is collected from hospital and death records then aggregated to the population level. Individual-level information is generally not publicly available for confidentiality reasons, but can be made available to researchers through the use of administrative safe havens (subject to a successful application), such as the NHS Scotland National Safe Haven (http://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/#NSS-National-Safe-Haven). Disease data are classified according to the International Classification of Diseases (ICD), which is primarily used to report mortality and hospitalisation data; a popular indicator for the health status of a population. The ICD is maintained by the World Health Organisation (WHO, http://www.who.int/en/), with versions 9 (World Health Organisation, 1975) and 10 (World Health Organisation, 1994) currently used in practice, and the eleventh version being due to be released in 2018.

Within the air pollution and health literature, a number of classifications have been used to represent health, with some studies looking for associations with all-causes of mortality (Jerrett et al., 2005c). While a positive and significant association was found for particulate matter, considering all causes of death may not be the best indicator, since it contains deaths not related to air pollution exposure, thus potentially generating a biased result. Therefore, many studies tend to focus on cause-specific disease, such as those due to circulatory and respiratory conditions (Scoggins et al., 2004) or cardio-respiratory conditions (Wang et al., 2009). These types of cause-specific conditions are important to consider because they are more likely to be related to the adverse effects of air pollutants. Respiratory conditions, such as asthma and chronic

obstructive pulmonary disease (COPD), are particularly of interest, as exposure to gaseous and particulate air pollutants is likely to aggravate the respiratory tract when travelling into the lungs (Bernstein et al., 2004). It was estimated that, in 2008, more than 23 million Americans suffered from asthma, while 13 million adults suffered from COPD (Pleis et al., 2009). Therefore, it is important for policy-makers to identify risk factors for respiratory diseases so that incidence and financial burden can be reduced. Furthermore, it is argued that exposure to air pollution acts as an additional stress in persons who already stuffer from morbidities, such as those relating to chronic conditions (Anderson, 2009; Anderson et al., 2003).

While these studies help shed light on the associations between air pollution and health, they do suffer from the major drawback of lacking statistical power to be able to detect associations. By focusing on specific causes of disease, one will reduce the number of disease cases utilised in the analysis. Willocks et al. (2012) discussed this phenomena in detail, where they explained that, in order for an association to be detected, there has to be sufficient variation in the pollution and disease data. In the case of their study where no association was observed, the levels of variation in the air pollution and disease data were low, and the inter-quartile ranges for the daily counts of cardiovascular hospital admissions and levels of PM_{10} (measured in μgm^{-3}) were between $(8, 13)$ and $(17, 30.5)$ respectively. The authors also suggested that utilising routinely collected data may not provide enough variation to allow an association to be detected, especially when the region under study is considered to be relatively small. However, their study was conducted at the daily level. Aggregating to a monthly or annual level may be sufficient to mitigate the issue of low variation in the response and covariates.

While seeking an association between specific causes of death and pollution is important, it is also of interest to ascertain whether there is a greater effect of air pollution in specific age groups. Numerous studies have observed stronger effects of air pollution on health in the elderly population (Fischer et al., 2003; Larrieu et al., 2007; O'Neill et al., 2004). It is believed that exposure to air pollution at these ages is more likely to cause harm or exacerbate cardiac and respiratory conditions in a population which is classed as more vulnerable compared to the general population. Just as the elderly are considered more susceptible to the adverse effects of air pollution, children are also considered as a vulnerable group with respect to air pollution (Beatty & Shimshack, 2014). Children are more likely to suffer from direct exposure to air pollution since they exhibit greater activity levels and spend more time outdoors, which can lead to variable breathing rates and thus affect lung function (Beatty & Shimshack, 2014). Furthermore, it is widely acknowledged that children are more prone to suffering from chronic respiratory conditions, such as asthma, meaning that exposure to air pollution can aggravate and exacerbate these chronic conditions. In addition, the effect of air

pollution on infant mortality (Padilla et al., 2013), preterm delivery (Yi et al., 2010), and preterm birth (Johnson et al., 2016) has also been investigated. The majority of this thesis focuses on mortality and morbidity due to cardio-respiratory diseases in the adult population (Chapters 5 and 6). However, this thesis also investigates the association between air pollution and ill health at different ages (Chapter 6), where age is stratified into 3 age groups: younger population (0-19 years); working population (20-64 years); and an older population (65+ years). This allows for a deeper understanding as to which ages air pollution has a more detrimental effect. This has important policy implications when trying to find the best ways in which to target air quality interventions.

3.3.3 Air pollution data

The term 'air pollution' includes a wide variety of atmospheric pollutants that are present as gases or particles, which are individually and routinely monitored at hourly intervals by a network of outdoor monitoring stations. Due to high correlations between individual pollutants, the majority of pollution-health studies estimate health effects based on exposure to a single pollutant, or use multiple pollutants in separate models. Commonly-used measures of gaseous pollutants include carbon monoxide (CO, Villeneuve et al., 2003), oxides of nitrogen (Bennett et al., 2014), ozone (O_3, Tao et al., 2012), and sulphur dioxide (SO_2, Wong et al., 2008). Particle pollutants comprise black smoke (BS, Beverland et al., 2014a), and particulate matter, the latter characterised according to its aerodynamic diameter as either less than 2.5 μgm^{-3} ($PM_{2.5}$, Cesaroni et al., 2013 or less than 10 μgm^{-3} (PM_{10}, Pirani et al., 2014). $PM_{2.5}$ is considered one of the most important pollutants because it is small and can be easily inhaled into the lungs, and is therefore an important risk factor for lung cancer, respiratory and cardiopulmonary mortality (Pope III et al., 2002). It has been argued that studies should focus more on ultrafine particles (aerodynamic diameter less than 0.1 μgm^{-3}) as these are what makes up the majority of the particle pollutants in urban and industrial areas (Terzano et al., 2010). However, in the UK, these smaller particles are not monitored making it difficult to assess their health effects.

In this thesis, nitrogen dioxide (NO_2, Huang et al., 2015) is the main focus, as it is a good marker for traffic-related air pollution, produced from vehicle exhausts (World Health Organisation, 2006), and also due to its strong correlation with other traffic-related pollutants (Brunekreef & Holgate, 2002). Furthermore, data on NO_2 concentrations were more widely available in terms of the number of spatial locations of the monitoring network compared to particulate matter, meaning it is a more comprehensive measure of air pollution across West Central Scotland.

3.3.4 Covariate data

In addition to air pollution, a wide variety of covariates are used to model risk factors that are related to the disease data. If these covariates are not included in the model, spatial autocorrelation and overdispersion can be induced, and thus any resulting pollution-health relationship could be biased. This is known as confounding, where an important spatially-correlated variable is not included in the model and the resulting pollution-health risk is stronger or weaker than what it would have been had if the variable had actually been included.

In time series studies, typical covariates include measures of meteorology, such as temperature (Beverland et al., 2014b; Ou et al., 2008) and humidity (Samoli et al., 2006), influenza epidemics (Hoek et al., 2000; Thach et al., 2010; Touloumi et al., 2005), measures of deprivation (Carder et al., 2010; Jerrett et al., 2004), and time-related covariates such as day of the week (Neuberger et al., 2013; Simpson et al., 2005). Cohort studies and spatial ecological studies mainly focus on aspects of deprivation, since it is an important determinant of health and environmental justice concerns have been raised with regards to differences in the effects of air pollution exposure being different across socio-economic groups (Laurent et al., 2007; O'Neill et al., 2003; Pellow, 2000), with more deprived populations being more strongly affected (Forastiere et al., 2007; Jerrett et al., 2004). This is argued to be due to populations living in more deprived areas having less adequate access to healthcare, poorer nutrition, lack of material resources, and a higher prevalence of smoking, which makes these populations more susceptible to the effects of air pollution compared to those populations residing in more affluent areas (Barceló et al., 2009; O'Neill et al., 2003; Richardson et al., 2011; Wong et al., 2008). These factors also make their health worse, irrespective of air pollution. Those residing in more affluent areas have a greater chance of living away from undesirable areas containing high traffic density or industrial facilities, due to being less financially constrained (Crouse et al., 2009).

Scotland is known for having the lowest life expectancy and highest mortality rates in Western Europe (McCartney et al., 2012; Schofield et al., 2016), with Glasgow containing more than 40% of Scotland's most deprived areas comprising at least half of Glasgow's total population (National statistics. Scottish Index of Multiple Deprivation, 2012). It is therefore important to take deprivation into account when assessing the pollution-health relationship in West Central Scotland, since deprivation is a known determinant of health. Figure 3.1 displays the potential pathways through which socio-economic position can increase exposure to air pollution, as well as susceptibility, which relates to the presence of pre-existing medical conditions that can make individuals more susceptible to the harmful effects of air pollution.

Deprivation, by nature, is multifactorial and studies have to account for depriva-

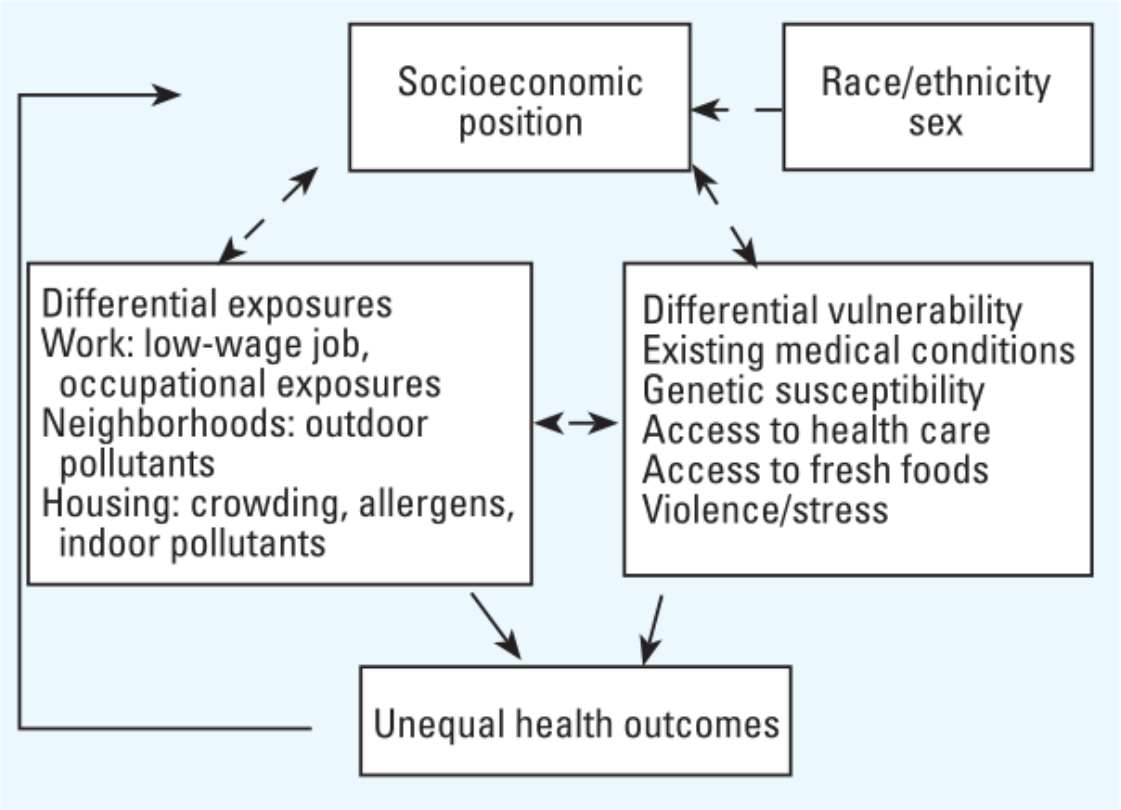

Figure 3.1: *Suggested pathways for socio-economic position to increase exposure and susceptibility to air pollution (taken from O'Neill et al. (2003)).*

tion through a number of proxy measures, such as the level of income or education within each areal unit. There have been numerous studies which have found significant associations between air pollution and ill health while adjusting for indicators of deprivation. For example, Wong et al. (2008) utilised information on education, income and unemployment to classify each area as being low, medium or highly deprived. Significant associations were found for numerous pollutants, including NO_2 and PM_{10}, with cause-specific mortality for middle and high deprivation areas, and the authors concluded that more deprived neighbourhoods experience increased mortality risks associated with higher levels of pollution. However, the deprivation index used in that study may not have captured the true extent of social deprivation in these areas, since only certain factors were taken into account, such as unemployment, and monthly household income. This could lead to an overestimation (or underestimation) of the observed pollution-health relationships. A review paper by Laurent et al. (2007) examined the literature (up until the year 2006) on the effect of socio-economic deprivation on the relationship between air pollution and mortality. The authors concluded that results were inconsistent across the studies, indicating that it was not possible to assert whether socio-economic deprivation modified the pollution-health relationship. However, they did conclude that studies measuring socio-economic deprivation at the individual level (i.e., in cohort studies) typically found that deprivation did modify the effect of air pollution on health, with deprived people having greater exposure to

air pollution. Studies that used a coarser measurement of deprivation, i.e., at the area level, either found inconsistent results or no statistically significant associations. While the majority of these studies did not adopt an areal unit design, it is interesting to note that in the studies of a more ecological nature, the relationship between air pollution and health was attenuated.

In areal unit studies, where inference is made at the ecological level, the efficacy of the deprivation measures utilised depends on the size of the areal units under study, as smaller units should be more socially homogeneous with respect to the level of deprivation (Leyland et al., 2007b). Higher risks of mortality are observed in areas with higher levels of pollution; however, due to the design of ecological studies, any individual-level measurements of deprivation cannot be taken into account, and so any observed associations may be over- or underestimated. There have been a number of studies which have utilised individual-level measurements of deprivation along with area-based measurements in order to examine the presence of residual confounding by the exclusion of individual-based measurements. Goodman et al. (2011); Jerrett et al. (2005b); Naess et al. (2007) all included individual- and area-level measurements of deprivation in their analyses. Jerrett et al. (2005b) controlled for 44 individual covariates, including education, income, and unemployment. Their results indicated that including the individual-level covariates attenuated the effect of air pollution on mortality by 5.6%. Including the ecological variables only reduced the pollution-health effect by 5%, but the resulting relationship was no longer significant at the 5% level. Similar results are seen in Naess et al. (2007). The authors compared models that included individual-level deprivation only with models including area-level deprivation only, and also to a fully adjusted model comprising both. All of the pollution-health relationships either stayed the same or were very similar across the models. For example, an indicator of primary education resulted in an attenuation of 4.5% when comparing the individual-level and area-level model, and an attenuated effect size of 3.6% for the fully adjusted model. Goodman et al. (2011) also found that adjusting for individual-level deprivation in addition to areal-level measurements did not explain any more of the association between air pollution and ill health. These results indicate that incorporating individual-level deprivation does not make a notable difference to the estimated pollution-health relationship. Studies that have not been able to include individual-level measurements may not be overestimating the observed findings and are unlikely to be affected by residual confounding of individual-level socio-economic deprivation. Adjusting only for area-based measures of deprivation appears to be sufficient when modelling the air pollution-health relationship.

Deprivation indexes are another way of tackling the multifactorial nature of deprivation, by combining a number of different socio-economic variables into one overall index. There are numerous area-level deprivation measures that have been used in

studies, such as the Carstairs score (Carstairs, 2001; Elliott et al., 2007), the Townsend Index (Haining et al., 2010; Maheswaran et al., 2005a, 2006, 2012; Townsend et al., 1988; Walters et al., 1995), the English Indices of Deprivation (Bennett et al., 2014; Maheswaran et al., 2012; Tonne et al., 2008, 2010), and the New Zealand Deprivation Index (Richardson et al., 2011; Scoggins et al., 2004). Both the Carstairs score and Townsend Index comprise a combination of census output variables, such as social class, overcrowding, unemployment, and car ownership; while the English Index of Multiple Deprivation and the New Zealand Deprivation Index are a combination of multiple domains relating to income, education and employment, among others. The variables and domains used to create these deprivation indexes are based on health information that is routinely collected from government departments, such as the Department for Work and Pensions, the Information Services Division (ISD), and the National Health Service (NHS). Other studies tend to combine a multitude of individual socio-economic variables available from census data into one overall index in order to accurately characterise small area deprivation levels (Havard et al., 2009; Jerrett et al., 2005c; Laurent et al., 2008; Padilla et al., 2013; Wong et al., 2008).

In Scotland, most studies make use of proxy measures of socio-economic deprivation, such as the median property price in each areal unit, and the proportion of people in each areal unit claiming job seekers allowance (Huang et al., 2015; Lee & Mitchell, 2012, 2014; Lee et al., 2014). However, Scotland does have its own Scottish Index of Multiple Deprivation (SIMD) that aims to establish a relative index of socio-economic deprivation across data zones. The SIMD is similar to the aforementioned area-based indexes in the sense that it comprises a number of different domains, including income, education, and employment, among others. While it is important for studies to account for deprivation in their analyses, it is unclear whether the relationship between air pollution and ill health is dependent on the choice of deprivation measure. Therefore, Chapter 5 presents a sensitivity analysis to the choice of deprivation measure on the estimated pollution-health relationship in West Central Scotland to establish whether the relationships between air pollution and ill health changes depending on the measure of deprivation used.

3.3.5 Standard spatial model

As previously mentioned, the disease, air pollution, and covariate data are available at the ecological level, where the disease data are in the form of counts. Count data are discrete and take the form of natural numbers $\{0, 1, 2, \dots\}$, which are whole and non-negative and are therefore assumed to arise from a Poisson distribution. These data denoted by $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ for all m areal units, and are regressed against air pollution concentrations $\mathbf{x} = (x_1, \dots, x_m)^\top$, and a matrix of p covariates denoted by $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_m)^\top$, which includes a column of ones for the intercept term. Typically, a generalised Poisson log-linear model (GLM) is used

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i R_i) \quad \text{for } i = 1, \ldots, m \\ \ln(R_i) &= \mathbf{x}_i \beta + \mathbf{z}_i^\top \boldsymbol{\delta} + \phi_i, \end{aligned} \tag{3.1}$$

where E_i represents the expected number of disease cases in areal unit i and is treated as an offset to allow for the number of disease cases to vary according to the age-sex structure and size of the underlying population. The risk of disease in areal unit i is denoted by R_i, and $(\beta, \boldsymbol{\delta})$ denotes the regression parameters for the air pollution and covariates respectively. This type of GLM (where $\phi_i = 0$ for all areal units i) is used as an exploratory technique to highlight the need for a spatial model by showing overdispersion is present, and by assessing the presence of residual spatial autocorrelation in the data, where Moran's I (see equation(2.76)) is calculated on the residuals.

If residual spatial autocorrelation is not present after adjusting for the covariates, then one need not progress onto a more complex modelling scheme. However, in most cases, solely including the covariates is not enough to account for the variation in the disease data. This then leads on to the inclusion of a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$ into the linear predictor of equation (3.1) as a way of capturing the leftover autocorrelation in the data. A number of models can be specified for these random effects, including conditional autoregressive (CAR), simultaneous autoregressive (SAR), or geostatistical models. However, in the spatial areal unit study context, CAR models (Besag et al., 1991) models are most common (Lawson et al., 2012; Lee et al., 2009; Lee & Sarran, 2015; Maheswaran et al., 2005a), and are discussed in Chapter 2.

A number of globally smooth CAR models have been developed, and a review by Lee (2011) concluded that the specification proposed by Leroux et al. (1999) was the most appealing, since it can represent a range of strong and weak spatial autocorrelation all within one set of parameters, unlike other CAR models. This modelling framework is typically implemented within a Bayesian setting, with inference based on McMC simulation; however, these models can also be implemented within a frequentist setting using maximum likelihood methods. Bayesian methods are becoming the norm in this context due to increasing improvements in computational power, making Bayesian methods quicker and easier to implement. The estimated pollution-health effect is typically reported on the relative risk (RR) scale, which measures the magnitude of increasing air pollution levels on the health of the population. The RR is reported in terms of a specific increase, usually a one standard deviation, in the pollutant, and is given by

$$\text{RR}(\beta) = \exp(\omega \times \beta), \tag{3.2}$$

where ω represents the standard deviation of the pollutant under study, and is used to ensure a realistic change in long-term exposure. For example, a RR of 1.2 corresponds to a 20% increased risk of disease.

3.4 Geographical locations

It is known that the relationship between air pollution and ill health has been well studied for the past two decades, and this section aims to provide an overview of the literature regarding areal unit studies conducted in Scotland, and in the wider UK, European and worldwide contexts. The main difficulty when comparing areal unit studies both within and across countries is the varying sizes of administrative boundaries, which can make it difficult to compare different pollution-health effects since they do not all relate to the same sized areas. Therefore, in this thesis, studies have been grouped according to their average areal unit population size, in order to attempt consistency when presenting the results.

The majority of studies conducted in Scotland (especially earlier studies) sought to quantify the relationship between air pollution and ill health through the implementation of cohort or time series designs (Agius et al., 2002; Beverland et al., 2014a, 2012a,b; Carder et al., 2010, 2008; Dibben & Clemens, 2015; Prescott, 2000; Prescott et al., 1998; Willocks et al., 2012; Yap et al., 2012), where no consistent associations were found between exposure to air pollution and ill health. However, the studies in which an ecological areal unit design was adopted (Huang et al., 2015; Lee, 2012; Lee et al., 2009; Lee & Mitchell, 2014; Lee et al., 2014) have collectively found substantial pollution-health effects for NO_2 and PM_{10} at the intermediate geography level (IG, median population of 3956) in Glasgow. The first paper in this field by Lee et al. (2009), investigated the relationship between PM_{10} concentrations and hospital admissions due to respiratory disease in four of the largest cities in Scotland, namely: Aberdeen, Dundee, Edinburgh, and Glasgow. For a one standard deviation increase in PM_{10} concentrations, the authors found a 4%, 5%, 7%, and 7% increase in respiratory hospital admissions for each city respectively. The results were substantial at the 5% level for Edinburgh and Glasgow as their 95% credible intervals did not contain the null risk of one; however, the same was not found for Aberdeen and Dundee. These results suggest that, in Edinburgh and Glasgow, areas with higher pollution levels have increased risk of respiratory hospital admissions. Similar results were reported in Lee (2012); Lee & Mitchell (2014); Lee et al. (2014), where the effect sizes ranged between 4% and 6.6%. Huang et al. (2015) studied Scotland as a whole and reported a 2.3% increase in respiratory hospital admissions for a one standard deviation increase in NO_2 concentrations.

Even within the wider UK context, there are relatively few studies that explore the relationship between air pollution and ill health using an ecological small area design.

Maheswaran et al. (2005a,b, 2006) utilised a small area design at the census enumeration district level in Sheffield comprising, on average, 194 inhabitants over the age of 45 years. Maheswaran et al. (2005a, 2006) observed a positive relationship between modelled NO_x concentrations and stroke mortality, where the highest risks were found in the highest NO_x category. However, Maheswaran et al. (2005b), did not observe any substantial relationships between NO_x, PM_{10} or CO and ill health due to coronary heart disease. Haining et al. (2010) extended the study by Maheswaran et al. (2006), and reported that higher levels of NO_x were associated with an increased risk in stroke mortality. This study used a more robust modelling framework to try and account for ecological bias (see Section 3.5); however, the authors concluded that the association observed by Maheswaran et al. (2006) (where the prospect of ecological bias was not considered) was not affected by ecological bias and is therefore a robust finding. Elliott et al. (2007) considered larger electoral wards (comprising on average 5300 inhabitants) for their unit of analysis, where stronger effects were observed between respiratory disease and black smoke (BS) and SO_2 compared to lung cancer and others. The effects were reported on the excess risk scale, where excess risks of 19.3% and 21.7% were observed for BS and SO_2 respectively, for an increase in 10 units for both pollutants.

A study by Tonne et al. (2008) evaluated the impact of the London congestion charge at the census ward level (comprising 1500 inhabitants on average) to the surrounding pollution levels (NO_2 and PM_{10}) and life expectancy. The authors observed a notable decrease in pollution concentrations as a result of the intervention, and concluded that more deprived areas had the highest levels of pollution. In addition, the London congestion charge scheme resulted in a small increase in life expectancy due to the reduction in traffic-related pollution, and therefore has an important public health impact. Tonne et al. (2010) also investigated whether there was an association between cardio-respiratory hospital admissions and NO_x concentrations across Greater London, also at the census ward level. However, no consistent associations were observed. In comparison, Maheswaran et al. (2012) sought to quantify the relationship between NO_2 and PM_{10} with the incidence of stroke, within smaller areal units comprising on average 283 inhabitants. No substantial associations were observed, but the authors did conclude that there was evidence of an increased risk among older-aged people between 65 and 79 years. The more recent studies conducted in England have focussed on areal units on a larger scale. Both Bennett et al. (2014) and Rushworth et al. (2014) studied the relationship between air pollution and ill health at the ward level, comprising 5000 inhabitants on average. These are a similar size to the areal units utilised in the studies conducted in Scotland. Bennett et al. (2014) observed a positive association between NO_x and the risk of heart failure in Warwickshire, while Rushworth et al. (2014) found substantial relationships in London between CO and $PM_{2.5}$ and respiratory hospital admissions, but not for PM_{10} or NO_2. Finally, Lee & Sarran (2015) studied the relationship between emergency respiratory hospital admis-

sions and NO_2, PM_{10} and $PM_{2.5}$, across local unitary authorities (population between 50,000 and 500,000) in England, and concluded there was evidence of substantial relationships between air pollution and respiratory ill health (relative risks of 1.089, 1.032 and 1.013 respectively).

There have been relatively few small areal ecological studies conducted elsewhere in Europe. Both Barceló et al. (2009) and Laurent et al. (2008) used census boundaries comprising 4000 and 2000 inhabitants respectively, but found inconsistent results between PM_{10} and numerous causes of mortality, including cardio-respiratory diseases. Barceló et al. (2009) only found significant results in the metropolitan area of Barcelona, and only for men. There have also been few studies conducted in North America, where each study utilised a different sized areal unit. Jerrett et al. (2005c) utilised the Canadian census tract comprising 3000 inhabitants on average, whereas Jerrett et al. (2005b) utilised zip codes in Los Angeles comprising 35,000 people on average. Hu et al. (2008) utilised the Florida census tracts comprising 4000 people on average, while Lawson et al. (2012) utilised counties in Georgia, where populations ranged from 1500 to 900,000. It can be argued that the smaller the areal unit, the more homogeneous the health, air pollution and deprivation data will be. All studies concluded that more deprived areas exhibited higher air pollution levels and higher rates of ill health, with Hu et al. (2008), and Jerrett et al. (2005b,c) further reporting positive associations between air pollution and health. However, Lawson et al. (2012), observed a negative association between $PM_{2.5}$ and asthma-related illness and described the finding as *'slightly surprising and inconsistent with some air pollution-related time-series studies'*. It is indeed a surprising result considering the majority of studies assessing the impact of fine particulate matter either observe a positive or inconsistent result, but the authors did awknowledge the ecological nature of their finding, which could be an artefact of the modelling technique used. In addition to the relatively few studies conducted in North America in terms of areal unit studies, there have been relatively few studies conducted in Australia (Wang et al., 2009), China (Wong et al., 2008), and New Zealand (Richardson et al., 2011). All three studies utilised small areas, comprising between 3000 and 6000 people; however, only Wang et al. (2009) found no consistent associations between air pollution (NO_2, O_3, SO_2) and cardio-respiratory mortality.

3.5 Ecological bias

As mentioned in the previous sections, spatial ecological studies are increasingly being used to investigate the adverse effects of air pollution on ill health utilising population summaries of these data. Ideally, one would want to investigate the association between air pollution and ill health at the individual level as this is the only way of directly determining a causal relationship between air pollution and ill health. However, due

to the length of time it takes for cohort studies to be conducted, and the need for individual-level data in which confidentiality issues may arise, these types of studies are not always feasible.

There has been much discussion over the last two decades regarding ecological studies and their flaws; see Elliott & Savitz (2008); Greenland & Morgenstern (1989); Haining et al. (2010); Shaddick et al. (2013); Wakefield (2008); Wakefield & Salway (2001) for further details. With ecological studies, inferences cannot be made from the area level to the individual level as the areas studied may be highly heterogeneous, and any generalisation to individuals depends on everyone within the area being similar, which is highly unlikely. This is known as the *ecological fallacy* (Selvin, 1958), where one assumes that associations observed at the area level also hold at the individual level. While ecological studies are useful for estimating the pollution-health relationship due to their relative inexpense and easy data acquisition, results are still treated with caution (Haining et al., 2010).

A number of researchers have tried to mitigate ecological bias and estimate a causal link by taking a mixed approach, i.e., including individual-level data along with the ecological data (Elliott & Savitz, 2008; Haining et al., 2010; Wakefield, 2008; Wakefield & Shaddick, 2006). Haneuse & Wakefield (2007, 2008) developed an approach for case-control data that evaluates the conditional distribution of aggregated, ecological data given the individual-level data, implemented both within a frequentist (Haneuse & Wakefield, 2007) and Bayesian setting (Haneuse & Wakefield, 2008). Another approach by Jackson et al. (2008, 2006) corrects for ecological bias by combining exposure-response information at the individual-level and ecological-level data for the exposure. Two joint regression models are specified within a Bayesian framework, which let the individual data inform the exposure-response coefficient, thus correcting for ecological bias. These two types of design are considered in terms of either individual-level models, taking into account ecological data in order to improve statistical power, or as an ecological study taking into account individual-level data to mitigate against ecological bias (Jackson et al., 2006). While these approaches try and account for ecological bias, Wakefield & Shaddick (2006) average the individual level risks to try and avoid ecological bias, rather than estimating the relative risks from averaging exposure. Wakefield (2008) suggested supplementing the ecological data with individual-level information as in the previous approaches, but with regards to the covariate data used in this thesis, Section 3.3.4 considered including individual-level information along with the aggregated pollution data and it did not seem to affect the overall estimated pollution-health relationship. Within the Scottish context, only one study investigated the effects of ecological bias on their results. The study by Lee et al. (2009) assessed the effects of pure specification bias (where the risk model at the individual level is nonlinear and changes upon aggregation), by adjusting the

regression model according to the parametric approach by Richardson et al. (1987). The authors found only very minor changes in the estimated pollution-health effect, thus concluding a weak effect of pure specification bias in their study.

3.6 Estimating exposure to air pollution

Since spatial ecological studies make use of routinely available health data, exposure to air pollution also must be available at the ecological level. However, exposure then relates to the average level of pollution experienced by the population as a whole, which is typically measured by an outdoor pollution monitoring network, whereby researchers take an average of the monitors' values within the region of interest and ascribe that population with a level of exposure. This can be deemed as a poor representation of exposure since the majority of the population migrate between areas, and between indoor and outdoor environments. Furthermore, the monitoring network does not take into account indoor pollution sources, for example from gas cookers and fireplaces. Indoor pollution exposure may also have an effect on the health of the population under study, which may vary across different age groups. For example, children spend a lot of their time outdoors and at school, which can be in a different area to which they reside in. The working age population (20-64 years) spend the majority of their time at work and are more likely to travel between different areas. People in the older age bracket ($\geqslant$ 65 years) may spend the majority of their time indoors, and be more likely to stay within their residential neighbourhood. In addition, averaging concentrations from monitors ignores any small-scale variation in the pollution concentrations, since the measurements from the monitors are realisations of the actual pollution field, which is spatially varying. Therefore, there is a need for more appropriate estimates of exposure at the population level to be used in spatial ecological studies.

There is major concern in the air pollution and health literature of obtaining air pollution data that are comprehensive and of good quality. Generally, information on air pollution is available from two distinct sources: the aforementioned outdoor monitoring stations, and modelled concentrations from numerical models. The data from the monitoring network are at the point level since these are observations taken from monitoring sites located throughout the study area, while modelled concentrations are estimated over a regular grid, such as 1km intervals. One major drawback of solely utilising the observed concentrations from the monitoring network is that the network is sparse across the study area, meaning that not all areal units contain an air pollution value. The monitoring sites are also likely to be preferentially located where the pollution is thought to be highest (Zidek et al., 2014) and exceeds EU standards. This can produce biased estimates of the true pollution concentrations in terms of inflating area-wide concentrations, while potentially biasing the resulting health risk. Furthermore, the monitoring network has a multitude of missing data that arise from monitors

becoming faulty or being relocated, which can therefore impede its functionality. Nevertheless, as the pollution concentrations are directly measured they provide close to the true value with little measurement error (Gelfand & Sahu, 2010).

Numerical computer models are increasingly being used to estimate pollutant concentrations in the atmosphere, which mathematically estimate the underlying physical and chemical processes of the environment using partial differential equations. These numerical computer models require large amounts of data input, such as information on meteorological processes, land use, vehicle and power plant emission sources, which makes them computationally expensive and time-consuming to implement. Furthermore, the resulting outputs from these complex systems is often biased and does not hold any information about uncertainty in their estimates. The outputs from these numerical models is often calibrated against observations from the monitoring network (Pirani et al., 2014); however, the two types of data are on different spatial scales and thus not directly comparable. This is known as the *change-of-support* problem (Gelfand et al., 2001; Gotway & Young, 2002) in the statistical literature, that is, *'the problem of inferring about a spatial variable at a certain resolution using data with different spatial support'* (Berrocal et al., 2010b). This makes it problematic when investigating the relationship between air pollution and ill health since the health data are available as aggregated counts over irregular spatial units, while the pollution data from monitoring sites are at the point level and the numerical output is at the grid level.

There has been recent research interest in fusion modelling, which combines both types of air pollution data in order to produce spatially representative pollution concentrations that can be aligned with disease data. There are two main approaches to this. The first represents the true (unobserved) pollution field as an underlying latent process, which drives both the observed and modelled data (see Figure 3.2). The second approach is the regression-type method, which links the two data sources together.

3.6.1 Latent process-type approach

A latent process or variable is one that is not directly observed or measured, but is assumed to be related to variables that have been measured, such as the observations from the monitoring sites and the modelled output. This method assumes that both the monitoring and modelled data provide good information about the same underlying process, where each have their own error structure. The observed data are related to the true underlying process by a measurement error model, while the modelled data are related to the true underlying process by a linear model that accounts for the bias in the numerical estimates.

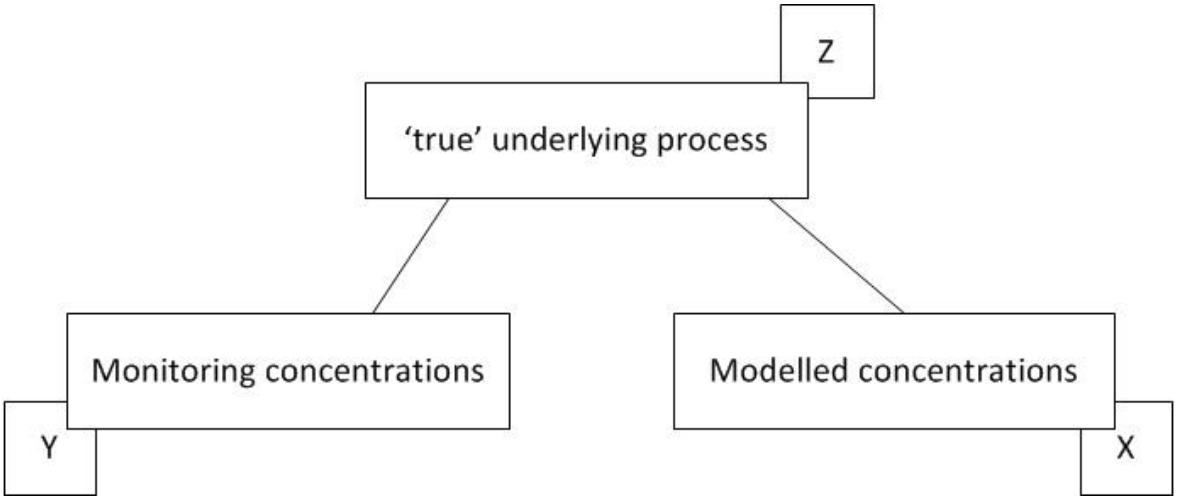

Figure 3.2: *Diagram of the latent process and its two components.*

Fuentes & Raftery (2005) adopted this approach to model the true environmental process at the point level, and is an application of the Bayesian melding method developed by Poole & Raftery (2000). In a purely spatial setting, let $Z(\mathbf{s})$ be the latent spatial process that measures the 'true' environmental factor at spatial location $\mathbf{s}$, which is assumed to follow the model

$$Z(\mathbf{s}) \;=\; \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{3.3}$$

where the spatial trend, $\mu(\mathbf{s})$, represents the overall mean of the process, and assumes that the latent process has zero-mean correlated errors $\epsilon(\mathbf{s})$. The monitoring site at location $\mathbf{s}$ is denoted by $Y(\mathbf{s})$, and is related to the latent process $Z(\mathbf{s})$, and measurement error $\delta(\mathbf{s})$, with the measurement error at location $\mathbf{s}$ distributed as a Gaussian process with mean zero and variance τ_Y^2. This model is of the form

$$Y(\mathbf{s}) \;=\; Z(\mathbf{s}) + \delta(\mathbf{s}), \qquad \delta(\mathbf{s}) \sim \mathrm{N}(0, \tau_Y^2). \tag{3.4}$$

The modelled data at location $\mathbf{s}$ are denoted by $X(\mathbf{s})$ and are modelled as

$$X(\mathbf{s}) \;=\; a(\mathbf{s}) + b(\mathbf{s})Z(\mathbf{s}) + \eta(\mathbf{s}), \qquad \eta(\mathbf{s}) \sim \mathrm{N}(0, \tau_X^2), \tag{3.5}$$

where the parameters $(a(\mathbf{s}), b(\mathbf{s}))$ control the additive and multiplicative bias of the modelled output respectively and are allowed to vary over space. However, because the modelled concentrations are averages over grid cells $(B_1, ..., B_c)$ that cover the entire study region D, the modelled concentrations are expressed in terms of stochastic integrals of each component of equation (3.5), that is

$$X(B_i) \;=\; \int_{B_i} a(\mathbf{s})\; d\mathbf{s} + \int_{B_i} b(\mathbf{s})Z(\mathbf{s})\; d\mathbf{s} + \int_{B_i} \eta(\mathbf{s})\; d\mathbf{s}, \tag{3.6}$$

for $i = 1, ..., c$. In this context, the bias in the modelled data is mostly additive, therefore the parameter $b(\mathbf{s})$ is treated as constant over space (i.e., $b(\mathbf{s}) = b$) in the

above equation (thus moving to the outside of the middle integral). These parameters are estimated by the following joint distribution between the monitored and modelled concentrations

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{a} + b\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_X \end{pmatrix} \right], \tag{3.7}$$

where $\boldsymbol{a}$ is the integral component from equation (3.6) ($\int_{B_i} a(\mathbf{s})$) evaluated at the c grid cells, and $\boldsymbol{\mu}$ is the integral of $\mu(\mathbf{s})$ evaluated at each of the grid cells $(B_1, ..., B_c)$. The goal is to predict the 'true' process at an unmeasured point location ($\mathbf{s}^*$), given by the predictive distribution

$$p(Z(\mathbf{s}^*)|\mathbf{Y}, \mathbf{X}) = \int p(Z(\mathbf{s}^*)|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})\ d\boldsymbol{\theta}, \tag{3.8}$$

where $(\mathbf{Y}, \mathbf{X})$ are the monitored and modelled spatial processes respectively, and Φ is the set of model parameters to be estimated.

This method of fusing the two types of environmental data together has two main limitations. The first limitation is that this method is computationally demanding and in most cases, infeasible when the modelled data contain a large number of grid cells. This is due to the large number of numerical integrations that would have to be computed at each step of the McMC algorithm. The second limitation follows on from the first in the sense that extending to a space-time domain is also computationally infeasible due to the computational burden it already possesses when considering a spatial component.

These limitations can be overcome by an upscaling fusion model, proposed by McMillan et al. (2010), in which the underlying latent process is specified at the grid cell level rather than at the point level. This method avoids the use of stochastic integration and also allows a temporal component to be incorporated. In this case, the specification of the monitoring concentrations in (3.5) has not changed, whereas the bias in the modelled concentrations is represented by a linear model of the form

$$X(\mathbf{s}) = Z(\mathbf{s}) + \boldsymbol{\beta} D(\mathbf{s}) + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \sim \mathrm{N}(0, \tau_X^2), \tag{3.9}$$

where $D(\mathbf{s})$ is a vector of bias covariates and $\boldsymbol{\beta}$ is a vector of parameters to be estimated in the model. The bias in the modelled concentrations $D(\mathbf{s})$ is expressed in terms of quadratic B-splines (Eilers & Marx, 1996) of the form

$$\sum_{j=1}^{N^D} D_{ij}\beta_j, \tag{3.10}$$

where $j = 1, ..., N^D$ is the total number of knots and i represents the bias covariates

for grid cell i. The latent process as defined in (3.3) can thus be extended to

$$Z(\mathbf{s},t) \;=\; \mu(\mathbf{s},t) + \epsilon(\mathbf{s},t), \tag{3.11}$$

where t represents time, and $\epsilon(\mathbf{s},t)$ are the spatially and temporally correlated errors modelled by a multivariate Gaussian distribution with an autoregressive (AR) prior of order one to model the temporal autocorrelation, and a conditional autoregressive (CAR) prior to model spatial autocorrelation. By modelling the latent process this way, it allows the model to be expressed in terms of all grid cells that can be indexed temporally and spatially. This model can predict on a much larger scale compared to previous models - daily concentrations for one year equating to over 9 million predictions. This approach also allows covariate data to be included in the model, but none was used in their study due to their covariate of choice having no predictive benefit.

Two additional papers utilise this latent process approach for obtaining spatially representative concentrations. Fuentes et al. (2008) follow on from their first approach at combining both sets of environmental data, but avoid the use of numerical integration by employing spatial logistic regression in order to model the probability of rainfall from monitored and modelled data at the grid cell level. The latent process is modelled as a Gaussian random field of the spatial and temporal domain with the time points modelled using a dynamic linear model of the form

$$Z(\mathbf{s},t) \;=\; \rho Z(\mathbf{s},t-1) + \beta W(\mathbf{s},t) + \epsilon(\mathbf{s},t), \tag{3.12}$$

where $\rho \in (0,1)$ controls the amount of temporal smoothing, and $\beta W(s,t)$ is a vector of weather covariates to improve the predictive performance of the model. Sahu et al. (2010) also identify a true underlying process with predictions at the point level. They developed a joint model by combining a space-time process for the monitoring data and a CAR model for the modelled data, then linked these two processes by using a latent space-time process in a Bayesian hierarchical modelling framework in order to avoid the spatial misalignment of the observed data. The main difference with this approach and the previous approaches is that two latent processes are chosen: one for the monitoring (point) data and one for the modelled (grid) data. The latent process for the monitoring data is assumed to follow the measurement error model (MEM)

$$H(\mathbf{s},t) \;\sim\; \mathrm{N}(V(\mathbf{s},t), \sigma_W^2), \tag{3.13}$$

where the term $V(\mathbf{s},t)$ denotes the latent process for the modelled data. Then, the latent process for the modelled data is assumed to follow a first-order AR model in time and a CAR model in space, that is

$$V(\mathbf{s},t) \;=\; \rho Z(\mathbf{s},t-1) + \eta(\mathbf{s},t), \qquad \text{for } i = 1,...,n, \quad t = 1,...,T, \tag{3.14}$$

where $\eta(s,t)$ are independent spatial processes, each of the form of improper CAR models applied at the grid cell level. The latent processes are introduced this way in order to capture point masses at zero with regards to chemical and wet deposition, while also avoiding stochastic integration. The monitoring data were considered as the ground truth, whereas the modelled data were expected to be biased. The measurement error model, therefore, allowed for the calibration of the modelled output.

3.6.2 Regression-type approach

The aforementioned limitations of the Fuentes & Raftery (2005) latent-process type approach can be overcome by fusing the monitored and modelled data together via a spatially varying linear regression as proposed by Berrocal et al. (2010b), which outperforms Fuentes & Raftery (2005) in terms of computational speed and out-of-sample validation (Gelfand & Sahu, 2010). In a purely spatial setting, let $Y(\mathbf{s})$ and $x(\mathbf{s})$ denote the monitored and modelled concentrations at spatial location $\mathbf{s}$. Berrocal et al. (2010b) propose a model of the form

$$Y(\mathbf{s}) \;=\; \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})x(\mathbf{s}) + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \sim \mathrm{N}(0, \tau^2), \tag{3.15}$$

where the monitored concentrations are assumed to be measured with little error, and are modelled as linearly-related to the error prone modelled concentrations. Due to the change-of-support problem, the modelled concentration nearest to each monitoring site is used in the above regression model to ensure that each monitoring site has a corresponding modelled grid cell value. Furthermore, a square root transformation of the monitored and modelled data was performed in order to stabilise the variance. The terms $(\beta_0(\mathbf{s}), \beta_1(\mathbf{s}))$ control the additive and multiplicative bias in the modelled pollution concentrations, and are allowed to vary over space. This spatial variation is captured via the coregionalisation prior (Gelfand et al., 2004), which is given by

$$\begin{pmatrix} \beta_0(\mathbf{s}) \\ \beta_1(\mathbf{s}) \end{pmatrix} \;=\; \mathbf{A} \begin{pmatrix} w_0(\mathbf{s}) \\ w_1(\mathbf{s}) \end{pmatrix}. \tag{3.16}$$

Here, $\mathbf{A}$ is a lower triangular matrix, while $(w_0(\mathbf{s}), w_1(\mathbf{s}))$ are two independent spatial processes. Specifically, each has a Gaussian distribution with a mean of zero and a correlation matrix defined by an exponential function of distance. The model was fitted in a Bayesian setting, using McMC simulation. This model is used to predict the monitored concentrations at locations without monitors using the above linear regression model, and measures of uncertainty in these predictions can be obtained from the McMC output via the posterior predictive distribution. That is, at iteration j of the McMC algorithm, the prediction at location $\mathbf{s}^*$ is given by

$$Y^{(j)}(\mathbf{s}^*) \;=\; \beta_0^{(j)}(\mathbf{s}^*) + \beta_1^{(j)}(\mathbf{s}^*)x(\mathbf{s}^*),$$

where $^{(j)}$ denotes the jth sample of a parameter from the McMC output, and $\left(\beta_0^{(j)}(\mathbf{s}^*), \beta_1^{(j)}(\mathbf{s}^*)\right)$ are obtained from Bayesian Kriging (as described in Section 2.4.1).

Another regression-type approach was proposed by Bruno et al. (2013), where they extended the previous regression approach by utilising a zero-inflated distribution to account for the high number of zero values in the monitoring data, considered as reliable measurements (in the context of rainfall). In a spatial and temporal setting, let $Y(\mathbf{s}, t)$ and $x(\mathbf{s}, t)$ denote the monitored and modelled concentrations at spatial location $\mathbf{s}$ and time t. A spatial logistic regression model is proposed, where the probability of monitor occurrence $\pi(\mathbf{s}, t)$ at spatial location $\mathbf{s}$ and time t is regressed by log-transformed modelled data of the form

$$\text{logit}(\pi(\mathbf{s}, t)) \;=\; \beta_0(t) + \beta_1(t)x(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \qquad \epsilon(\mathbf{s}, t) \sim \text{N}(\mathbf{0}, \sigma_{\epsilon t}^2 \mathbf{\Sigma}_\epsilon). \tag{3.17}$$

The variance parameter, $\sigma_{\epsilon t}^2$, contains a temporal component in order to capture the different variability in the probability of monitoring occurrence along time. The spatial adjustment, $\epsilon(\mathbf{s}, t)$, is specified as a multivariate Gaussian spatial process with mean zero and a correlation matrix $\mathbf{\Sigma}_\epsilon$ defined by an exponential covariance function of distance between sites $\mathbf{s}$ and $\mathbf{s}'$ and is of the form $\mathbf{\Sigma}_\epsilon = \exp(-\phi_\epsilon d_{\mathbf{ss}'})$ (where $d_{\mathbf{ss}'}$ donates the Euclidean distance between the sites). The bias terms $(\beta_0(t), \beta_1(t))$ were modelled as Gaussian distributions with mean zero and variance $\sigma_{\beta_0}^2$ and $\sigma_{\beta_1}^2$ respectively. The model was fitted in a Bayesian setting, adopting McMC simulation.

Considering the approaches by Berrocal et al. (2010b) and Bruno et al. (2013) to the spatial misalignment issue, Berrocal et al. (2010b) benefits from its simplicity and flexibility due to the modelling of the coefficients, and easily allowing for an extension to the temporal domain. It allows prediction at the point level; however, only Gaussian variables are acceptable in this context, restricting its application in other areas. Similarly, Bruno et al. (2013) easily allow a temporal component to be incorporated and extended (Berrocal et al., 2010b) by adopting generalised linear models, namely logistic regression, as their regression model of choice. Both methods exploit only the modelled data that correspond to the monitoring sites for parameter estimation (utilising all modelled data for prediction), but this does not restrict the quality of the predictions that can be produced at unmeasured locations. Both methods also do not take into account any other covariates (such as temperature) that might help increase the predictive performance of the models; however, this can be easily incorporated. Bruno et al. (2013) initially performed the modelling without taking into account the spatial aspect of the data, highlighting the need to include the spatial component in the modelling as the predictive performance of the model was notably improved. However, they did remark that reliable predictions outside of the monitoring network were hin-

dered, even when modelled data were available, due to the latent nature of the spatial effects. Nevertheless, both approaches adopt a simple, yet effective way of combining monitored and modelled data in order to provide a set of reliable measurements at the appropriate spatial scale.

3.6.3 Additional approaches

Studies by Huang et al. (2015); Pirani et al. (2014); Sacks et al. (2014); Vinikoor-Imler et al. (2013, 2014); Warren et al. (2013); Zhu et al. (2003) have adopted the fusion and latent process approaches described above to estimate air pollution concentrations and correct for spatial misalignment. Other common methods of estimating air pollution concentrations range from simple averaging of modelled concentrations to the correct spatial domain (Lee et al., 2009; Maheswaran et al., 2006; Rushworth et al., 2014; Warren et al., 2012), interpolation methods (involving inverse distance weighting and kriging (Elliott et al., 2007; Janes et al., 2007)) to the popular land-use regression (Bertazzon et al., 2015; Fernández-Somoano et al., 2013; Hansell et al., 2016), of which detailed reviews can be found in Jerrett et al. (2005a), and Hoek et al. (2008).

Briefly, land-use regression aims to predict pollutant concentrations by regressing observed concentrations from a small number of air pollution monitors from numerous other land-use covariates, such as distance to nearest road and meteorological variables. This method is also combined with a geographic information system (GIS) in order to model small scale variations in air pollution concentrations. Air dispersion models, which model air pollution concentrations based on emissions sources, are usually what modelled concentrations are based on, and it has been argued that dispersion models are more reliable compared to land-use regression models in intra-urban settings because they better represent the underlying process (Jerrett et al., 2005a). Beelen et al. (2010) provide a performance comparison between land-use regression models and dispersion models in predicting NO_2 concentrations in the Netherlands, and found a moderate agreement in the modelled concentrations between the two methods. In addition, measurement error models (MEMs) are increasingly being used to estimate pollution concentrations as part of a 2-stage modelling approach, whereby an exposure MEM is proposed at the first stage, where measurement error is characterised by the difference between the predictions and the 'true' unmeasured values, then these predictions are treated as a covariate and fed into a second-stage health model in order to estimate the possible exposure-health effect (Szpiro & Paciorek, 2013). However, the pollution model used in this thesis is based on the regression-type approach by Berrocal et al. (2010b); further details of which can be found in Chapter 4.

Chapter 4

Improving spatial nitrogen dioxide prediction using diffusion tubes: A case study in West Central Scotland

4.1 Introduction

As discussed in Chapter 3 Section 3.6, there is an inherent difficulty in obtaining air pollution data that are of good quality and will cover the entire study region. In the majority of air pollution and health studies, air pollution data typically come from a small number of automatic monitors that measure individual pollutants, such as NO_2 and PM_{10}, at a single point in space. However, the number of monitors is small and their geographical positioning is sparse, which does not allow an accurate representation of the spatial variation in air pollution concentrations that is required for epidemiological studies, particularly cohort and spatial ecological studies.

For cohort studies, concentrations are required at the residence of each member in the cohort, while for spatial ecological studies concentrations are required for each spatial unit at which health data are available. These fine scale pollution data are not available, for example, in the Glasgow area studied in this thesis, there are only 16 monitors covering the 368 square kilometre (km^2) study area. Therefore, inexpensive non-automatic diffusion tubes are also used to measure ambient concentrations of NO_2, and due to their lower cost and simpler equipment compared with automatic monitors, they are more prevalent. In the Glasgow study area considered here, there are 230 diffusion tubes, thus providing greatly enhanced spatial coverage compared with using the 16 automatic monitors alone.

The NO_2 data collected from the diffusion tubes are aimed at monitoring long-term exposure, usually monitoring at a monthly level, and then annual mean concentrations can be calculated. Conversely, the automatic monitors can accommodate a wide rage of exposure periods by recording hourly levels of multiple pollutant concentrations,

where daily, monthly and annual concentrations can be calculated.

However, combining these two data sets still does not give complete spatial coverage of the area under study, as is illustrated in the case study in Figure 4.2. As discussed in Chapter 3 Section 3.6 modelled concentrations from atmospheric dispersion models are used instead (for example, see Naess et al., 2007), as they provide estimated concentrations on a regular grid and thus have complete spatial coverage of the study region. However, these modelled concentrations are known to contain biases (Berrocal et al., 2010b), and are not as accurate as the measured pollution data.

Ideally, one would require measurements from automatic monitors and diffusion tubes at every possible location; however, only the spatially-dense gridded modelled data are available. NO_2 measurements can then be estimated at all gridded locations based on the modelled air pollution data and other relevant covariates. Therefore, this chapter proposes a geostatistical fusion model, that regresses the combined NO_2 concentrations from both automatic monitors and diffusion tubes against modelled NO_2 pollution data from an atmospheric dispersion model. This model is implemented within a Bayesian setting and predicts NO_2 concentrations across the Glasgow region for use in the pollution-health studies conducted in Chapter 5 and Chapter 6. This chapter demonstrates the dramatic improvement in fine scale spatial prediction of NO_2 that can be obtained by using abundant diffusion tube data that is relatively inexpensive to collect in addition to the small numbers of automatic monitors.

The remainder of this chapter is organised as follows. Section 4.2 describes the study design of the Glasgow case study, specifically the spatial extent of the region of interest, the NO_2 and covariate data. Section 4.3 presents the geostatistical fusion model for predicting NO_2 concentrations across the study region proposed here, and discusses its implementation. Section 4.4 presents the results of the analyses, including a model validation exercise that compares the proposed model against a number of other candidate models, and a fine scale prediction of NO_2 across the Glasgow region. Finally, Section 4.5 provides a concluding discussion.

4.2 Glasgow case study

4.2.1 Study region

The study region is centred around the Greater Glasgow conurbation, which is the largest city in Scotland, UK (see Figure 4.1). The Glasgow conurbation contains just under one quarter of the total Scottish population, equating to around 1.1 million people, with a land area of around 368 km^2. Seven local authorities comprise the study region, namely: East Dunbartonshire, East Renfrewshire, Glasgow City, North

Lanarkshire, Renfrewshire, South Lanarkshire, and West Dunbartonshire. These local authorities have been selected because they surround and include the city of Glasgow, collectively known as *West Central Scotland*, and include both urban and rural environments, which leads to a wide variation in pollution concentrations across the study region.

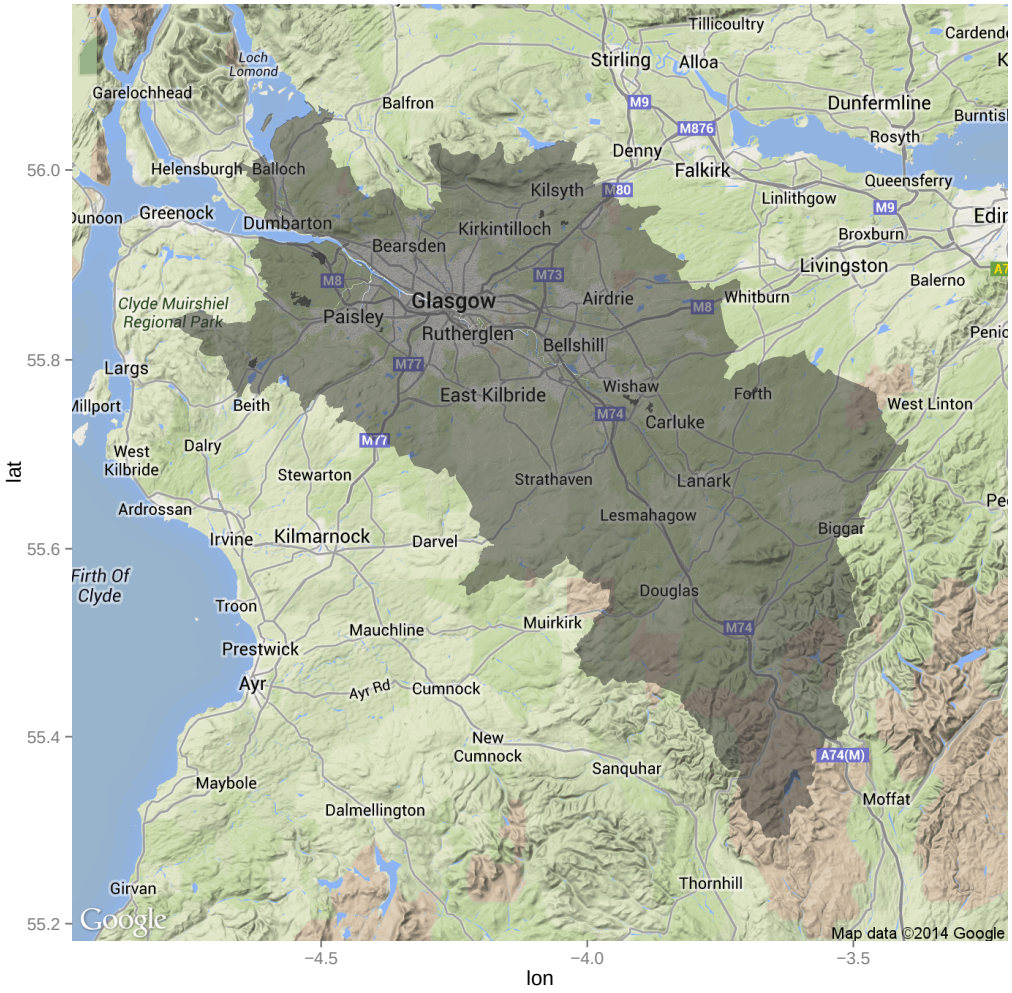

Figure 4.1: *Map displaying the chosen study region West Central Scotland.*

4.2.2 Air pollutant data

The air pollution data comprise annual mean concentrations of nitrogen dioxide (NO_2, measured in microgrammes per cubic metre μgm^{-3}) in 2006, for which two sources of data are available. The first source is measured data at fixed points in space, which come from both automatic monitoring stations and non-automatic diffusion tubes. NO_2 concentrations from the automatic monitors were downloaded from the Scottish Air Quality website (`www.scottishairquality.co.uk`), while the non-automatic diffusion tube data were obtained on request from each local authority through their Air Quality Progress Reports. These reports are required for each local authority in Scotland by the Local Air Quality Management (LAQM) process as set out in Part IV of the Environment Act (1995), the Air Quality Strategy for England, Wales and

Northern Ireland 2007, and the relevant Policy and Technical Guidance documents. This process ensures that all local authorities in Scotland regularly review, assess and report air quality within their areas, determining whether the air quality objective will be achieved. Air quality objectives in Scotland are determined by the Air Quality (Scotland) Regulations 2000 (Scottish SI 2000 No 97) and the Air Quality (Scotland) Amendment Regulations 2002 (Scottish SI 2002 No 297), which, for NO_2, state that an annual mean concentration of $40\mu gm^{-3}$ should not be exceeded, otherwise the site must undergo formal investigation.

The accuracy of the diffusion tubes varies depending on numerous factors, such as the preparation methodology used, handling procedures and the laboratory analysing the data. The diffusion tubes are calibrated using a bias-adjustment factor obtained from co-location studies between diffusion tubes and automatic monitoring stations. The data are collected and analysed by the local authority's chosen laboratory and the bias-adjustment factors are calculated through the methodology proposed by DEFRA (http://laqm.defra.gov.uk/bias-adjustment-factors/national-bias.html). In the Glasgow area all diffusion tubes were adjusted by a factor of 0.82. The NO_2 concentrations are measured at 246 sites across West Central Scotland, of which 230 are diffusion tubes and 16 are automatic monitors. The locations of these sites within the study region are displayed in Figure 4.2, where the diffusion tubes are displayed as crosses and the automatic monitors are presented as triangles. Summary statistics for the measured data are shown in Table 4.1. These statistics highlight that the distribution of NO_2 concentrations across West Central Scotland is slightly higher for automatic monitors compared to diffusion tubes, with median values of $34.55\mu gm^{-3}$ and $29.95\mu gm^{-3}$ respectively. This could be due to local authorities placing automatic monitors where they have a compliance problem with EU pollution standards. It could also be due to there being more roadside monitors (142 sites out of 246) placed along major roads throughout the study region, thus providing elevated NO_2 pollutant levels.

Table 4.1: *Summary statistics for the automatic monitoring and diffusion tube NO_2 (μgm^{-3}) data for 2006 across West Central Scotland.*

	Monitors	Diffusion tubes
Min	10.00	9.00
25th Percentile	29.35	22.25
Median	34.55	29.95
Mean	38.31	31.63
75th Percentile	42.50	38.00
Max	89.00	86.10

Figure 4.2 highlights the sparsity of the measured data, and shows that these automatic monitors and diffusion tubes do not provide complete spatial coverage of West

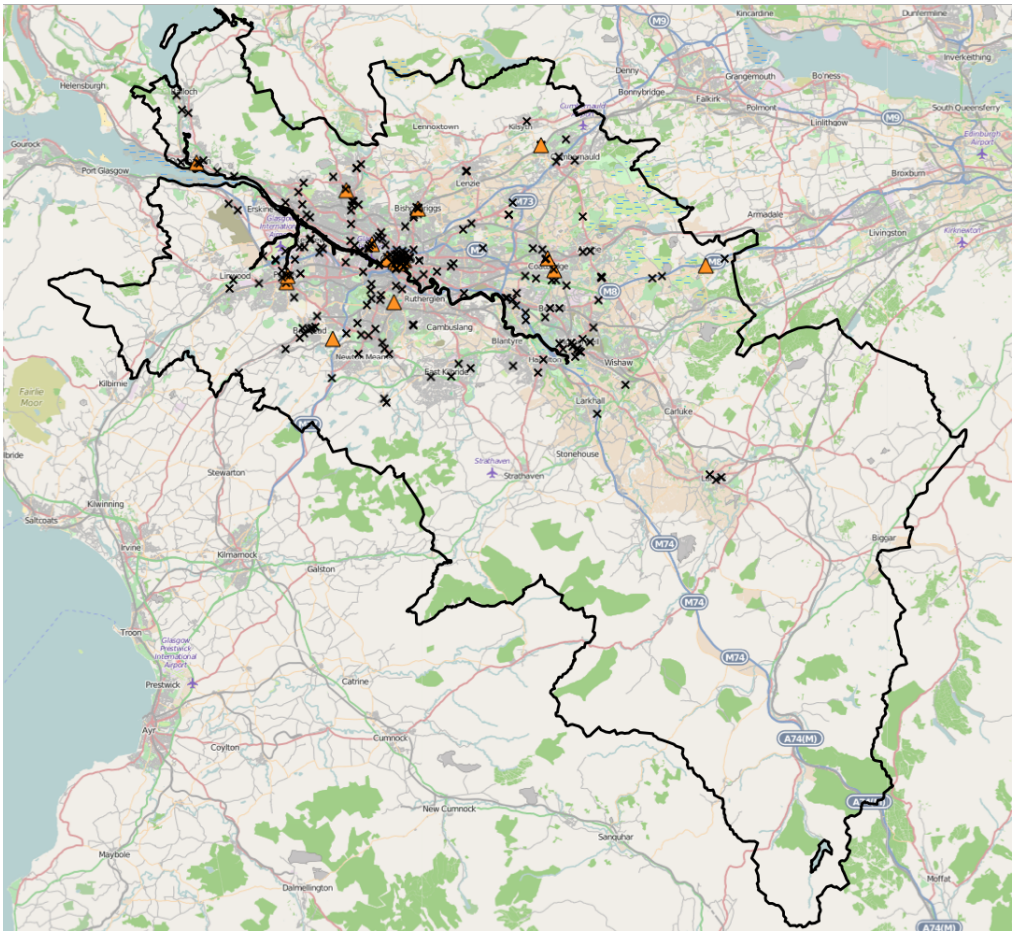

Figure 4.2: *This map showcases the locations of the measured (automatic monitors and diffusion tubes) NO_2 data for 2006 with the outline of the West Central Scotland study region. Crosses denote diffusion tubes, and triangles denote automatic monitors.*

Central Scotland.

The second source of data are modelled concentrations based on the UK Pollution Climate Mapping (PCM) approach (Brookes et al., 2011), provided by the Department for Environment, Food and Rural Affairs (DEFRA) (`http://uk-air.defra.gov.uk/`). These data are modelled as yearly mean background concentrations, measured in μgm^{-3}, at a 1km grid square resolution, thus providing complete spatial coverage across West Central Scotland with no missing data. However, modelled concentrations such as these are known to contain biases due to being uncalibrated, and no measure of variability is available to quantify the level of uncertainty in these estimates. These data are displayed in Figure 4.3, where the city of Glasgow and the main motorway network are easy to see. Annual mean concentrations range between 3.021μgm^{-3} and 34.760μgm^{-3}, with a mean value of 7.632μgm^{-3} across West Central Scotland. These concentrations are lower compared to the measured data as they are average background concentrations over a 1km square grid, rather than relating to specific pollution sources, such as roads.

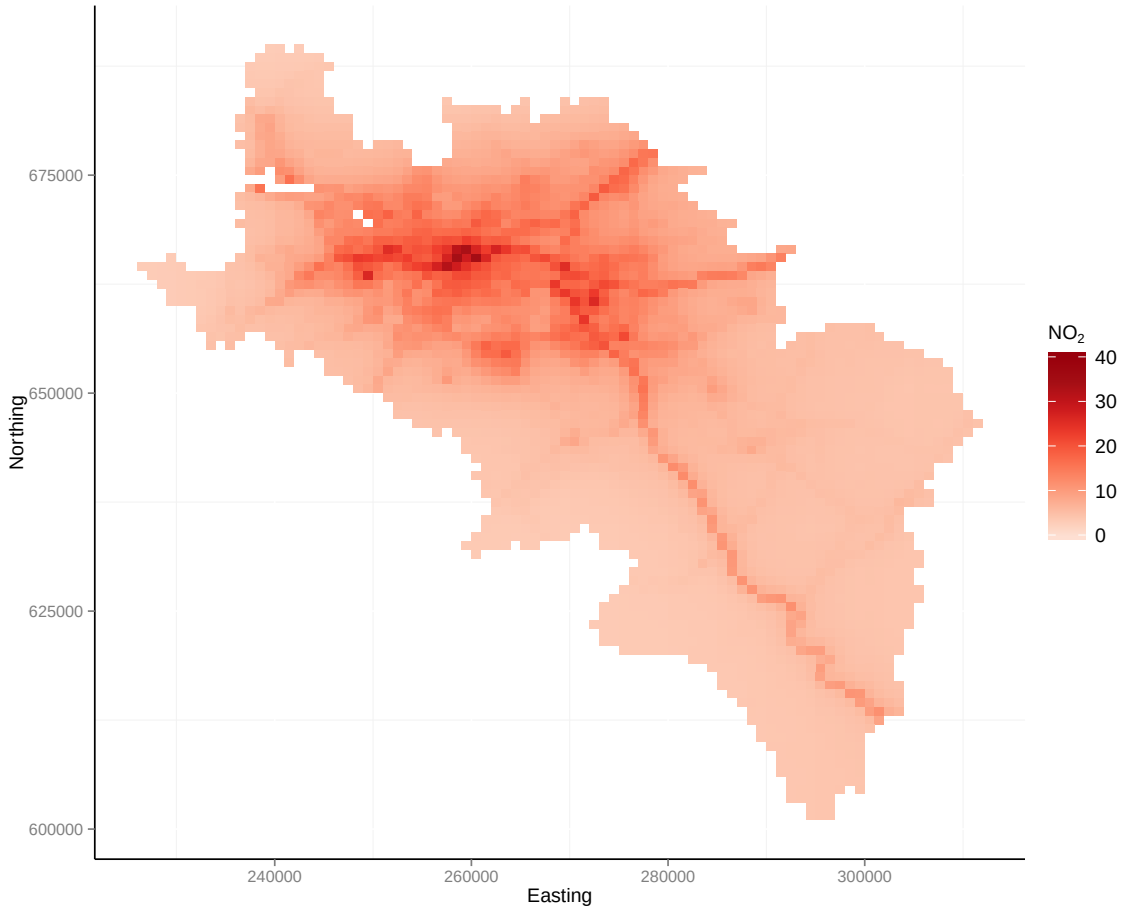

Figure 4.3: *Map showcasing the 2006 modelled NO_2 (μgm^{-3}) concentrations from an atmospheric dispersion model across West Central Scotland at a 1km grid square resolution.*

4.2.3 Covariate data

A number of covariates were considered in this study. Firstly, an indicator variable was included in order to distinguish the pollution concentrations measured from the two types of equipment: automatic monitors and diffusion tubes. Secondly, the local environment in which each of the automatic monitors and diffusion tubes were located was also recorded. These local environments include kerbside (placed within 1m of the kerb on a busy road), roadside (placed between 1m and 5m of a busy road), urban background (placed away from direct sources, usually in urban residential areas), rural (placed in countryside locations far from roads, populated and industrial areas), and special (placed at Glasgow airport and industrial sources). In total, there were 142 roadside, 34 kerbside, 8 special, 60 urban background, and 2 rural sites. Thirdly, in order to distinguish between urban and rural environments an urban-rural variable was constructed, which classifies each prediction location as urban or rural according to the Scottish Government 6 fold Urban Rural Classification (http://www.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification/) shown in Table 4.2. This classification is the primary framework for defining rural areas in Scotland. A location was considered urban if it was situated in a built-up area containing more

than 10,000 people (groups 1 and 2), and rural otherwise (groups 3-6).

Table 4.2: *Scottish Government 6 fold Urban Rural Classification.*

Classification	Description
1 Large urban areas	Settlements of $\geqslant$ 125,000 people.
2 Other urban areas	Settlements of 10,000-124,999 people.
3 Accessible small towns	Settlements of 3,000-9,999 people and be within a 30 minute drive of groups 1 and 2.
4 Remote small towns	Settlements of 3,000-9,999 people and be outwith a 30 minute drive of groups 1 and 2.
5 Accessible rural	Population $<$ 3,000 people and within a 30 minute drive of groups 1 and 2.
6 Remote rural	Population $<$ 3,000 people and be outwith a 30 minute drive of groups 1 and 2.

4.3 Statistical methods

This section presents the geostatistical fusion model proposed in this thesis for predicting NO_2 concentrations across West Central Scotland, using both the measured (automatic monitors and diffusion tubes) and modelled air pollutant data. Subsection 4.3.1 presents the statistical fusion model, while Subsection 4.3.2 outlines the prediction methodology. The model is fitted in a Bayesian setting, with inference based on Markov chain Monte Carlo (McMC) algorithms, which were written in the `R` statistical programming language (R Core Team, 2015).

4.3.1 Spatial fusion model

Let $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)^T)$ denote the vector of (natural) log-transformed NO_2 concentrations from both the automatic monitors and diffusion tubes at spatial locations $(\mathbf{s}_1, \ldots, \mathbf{s}_m)$, where the latter are measured as Eastings and Northings in metres. The NO_2 data are log-transformed because they are non-negative and skewed to the right, and exploratory analyses suggested that a log-transformation improved the fit of the resulting regression models (see Figure 4.4). These measured NO_2 concentrations are regressed against a matrix of p covariates denoted by $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1)^\top, \ldots, \mathbf{x}(\mathbf{s}_m)^\top)^\top$, where the values relating to spatial location $\mathbf{s}_i$ are denoted by $\mathbf{x}(\mathbf{s}_i)^\top = (x_0(\mathbf{s}_i), x_2(\mathbf{s}_i), \ldots, x_p(\mathbf{s}_i))$. This covariate matrix includes a column of ones for the intercept term, the (natural) log-transformed modelled concentrations, and other relevant covariates, such as the local environment in which the observation is located (e.g., roadside, urban background, etc). Thus, this model fuses the measured and modelled NO_2 pollutant data via a

linear regression relationship.

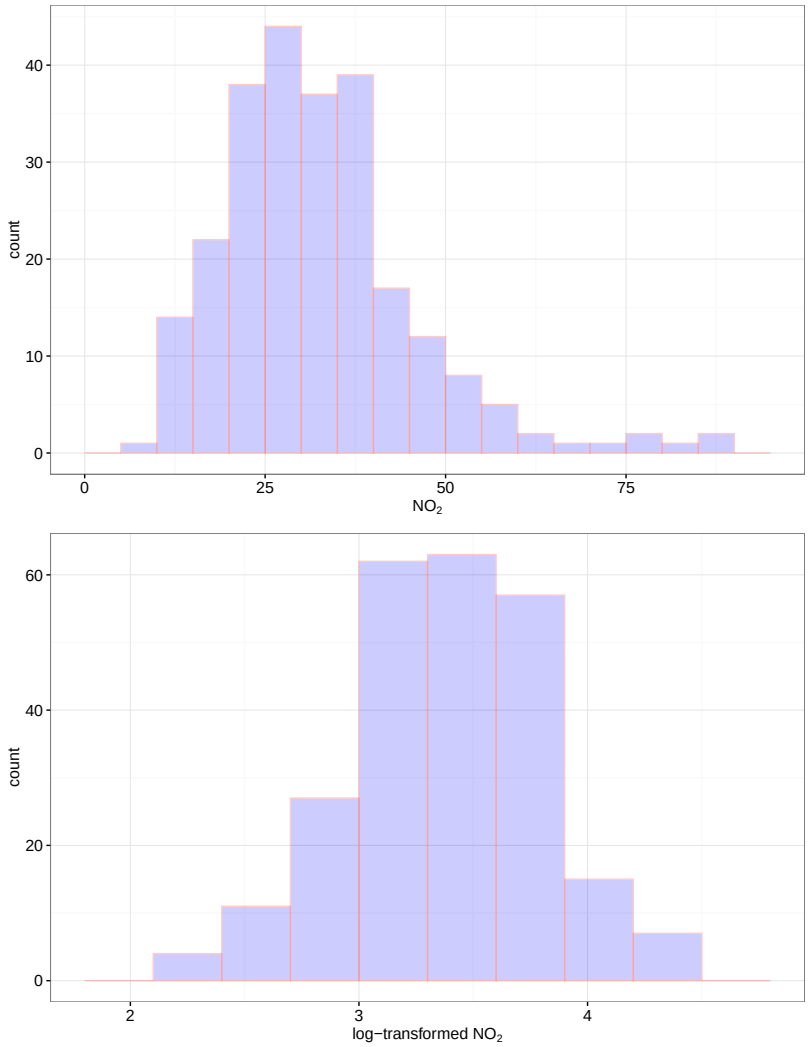

Figure 4.4: *The top panel displays the histogram for the NO_2 measured concentrations on the original scale, while the bottom panel displays the log-transformed measured NO_2 concentrations.*

A Bayesian geostatistical fusion model is proposed for these data, which relates the measured and modelled NO_2 concentrations using the equation (4.1). This model is a development of the model presented by Berrocal et al. (2010b) (discussed fully in Chapter 3 Section 3.6.2) in terms of making use of the diffusion tube data to increase the total number of observations and hence spatial locations. Furthermore, this newly developed specification includes other relevant spatial covariates to improve prediction, instead of utilising spatially varying coefficients, as in the Berrocal et al. (2010b) framework. These spatial covariates, as discussed in the previous section, are included to differentiate any differences in the spatial pattern of the observations, for example, in terms of differentiating between urban and rural environments. This specification here is simpler than the model developed by Berrocal et al. (2010b), but as will be shown

in further sections, produces a high prediction performance that rivals their model.

$$Z(\mathbf{s}_i) \sim \mathrm{N}\Big(\mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \phi(\mathbf{s}_i), \nu^2\sigma^2\Big), \quad i = 1, \ldots, m. \tag{4.1}$$

The mean function is a linear combination of a covariate component $\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ denotes the associated regression parameters for each covariate, and spatial random effect $\phi(\mathbf{s}_i)$. The regression parameters $\boldsymbol{\beta}$ are assigned a weakly informative multivariate Gaussian prior with mean zero and large diagonal variance matrix, such as $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(1000))$. The spatial random effects for all m locations are collectively denoted by $\boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_m))^\top$, and allow for any unmeasured spatial autocorrelation in the measured NO_2 data after the covariate effects have been accounted for. Their spatial autocorrelation is modelled using the formulation:

$$\begin{aligned} \boldsymbol{\phi} &\sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}(\rho)), \\ \sigma^2 &\sim \text{Inverse-Gamma}(a, b), \\ \rho &\sim \text{Discrete Uniform}(\rho_1, \ldots, \rho_r). \end{aligned} \tag{4.2}$$

The random effects $\boldsymbol{\phi}$ are assumed to come from a multivariate Gaussian distribution with mean zero, variance σ^2, and a spatial correlation matrix $\boldsymbol{\Sigma}(\rho)$. This matrix is defined by an isotropic exponential correlation function of the distance between any two locations, that is $\boldsymbol{\Sigma}(\rho) = \exp(-\rho\,\mathbf{D})$. Here $\mathbf{D}$ is an $m \times m$ distance matrix, where the ijth element, $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, is the Euclidean Distance between any pair of spatial locations $(\mathbf{s}_i, \mathbf{s}_j)$. In addition, the diagonal elements of $\mathbf{D}$ are zero corresponding to $d_{ii} = 0$. The exponential model was chosen for simplicity and because it is the most commonly used model in the geostatistical literature (see, for example, Vicedo-Cabrera et al., 2013). A conjugate inverse-gamma prior was specified for the spatial variance σ^2, where $(a = b = 0.001)$ were chosen to be non-informative. Here, ρ is the spatial decay parameter, which controls the rate at which the spatial autocorrelation between a pair of sites declines as the distance between them increases. A discrete uniform prior with a large range was specified for ρ as suggested by Diggle & Ribeiro (2007) for computational efficiency. This ensures that the correlation matrix, $\boldsymbol{\Sigma}(\rho)$, need only be inverted $r = 50$ times, once for each of the candidate values $\rho_1, \ldots, \rho_r$, rather than at every step of the McMC algorithm. Finally, the nugget effect, that is the amount of non-spatial variation or measurement error, is controlled by $\nu^2\sigma^2$, which is the product of the spatial variance parameter and the noise-to-signal ratio ν^2. This latter parameter is assigned a uniform prior on the unit interval, as the nugget effect is expected to be smaller than the amount of spatial variation for these data.

4.3.2 Spatial prediction

Bayesian spatial prediction using kriging is a natural extension in the Bayesian paradigm for estimating the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \nu^2, \sigma^2, \rho)$, and is implemented as a two-step procedure within the McMC algorithm. In models (4.1) and (4.2), spatial autocorrelation is induced into the mean function through the random effects $\boldsymbol{\phi}$. Therefore, the first step in spatial prediction generates the random effects at N prediction locations $\mathbf{s}^* = (\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*)$ using multivariate Gaussian theory as described fully in Chapter 2 Section 2.4.1.5. Specifically, the random effects at the prediction locations $\boldsymbol{\phi}^* = (\phi(\mathbf{s}_1^*), \ldots, \phi(\mathbf{s}_N^*))^\top$ are sampled from their conditional distribution given $\boldsymbol{\phi}$, that is

$$\boldsymbol{\phi}^* | \boldsymbol{\phi} \sim \mathrm{N}\left(\mathbb{E}[\boldsymbol{\phi}^* | \boldsymbol{\phi}], \mathrm{Var}[\boldsymbol{\phi}^* | \boldsymbol{\phi}]\right). \tag{4.3}$$

The mean and variance are given by

$$\mathbb{E}[\boldsymbol{\phi}^* | \boldsymbol{\phi}] = \mathbf{C}_Z(\mathbf{s}^*, \rho)^\top \boldsymbol{\Sigma}^*(\rho)^{-1} \boldsymbol{\phi}, \tag{4.4}$$

and

$$\mathrm{Var}[\boldsymbol{\phi}^* | \boldsymbol{\phi}] = \sigma^2 \left(\boldsymbol{\Sigma}^*(\rho) - \mathbf{C}_Z(\mathbf{s}^*, \rho)^\top \boldsymbol{\Sigma}^*(\rho)^{-1} \mathbf{C}_Z(\mathbf{s}^*, \rho) \right), \tag{4.5}$$

where $\boldsymbol{\Sigma}^*(\rho)$ is an $N \times N$ spatial correlation matrix for the N prediction locations and $\mathbf{C}_Z(\mathbf{s}^*, \rho)$ is an $N \times m$ spatial correlation matrix between the prediction and the observation locations. These equations are equivalent to ordinary kriging. The second step generates the predicted value of $Z(\mathbf{s}_i^*)$ for the N prediction locations as

$$Z(\mathbf{s}_i^*) \sim \mathrm{N}(\mathbf{x}(\mathbf{s}_i^*)^\top \boldsymbol{\beta} + \phi(\mathbf{s}_i^*), \sigma^2 \nu^2), \tag{4.6}$$

where $\mathbf{x}(\mathbf{s}_i^*)$ denotes the matrix of covariates at the N prediction locations.

Leave-one-out cross-validation is performed in order to assess the quality of the predictions, which removes each measured data point in turn and predicts its value from the remainder of the data. The accuracy of the predictions compared to the measured NO_2 concentrations are compared using three statistics, namely bias, root mean square prediction error (RMSPE), and the coverage probabilities of the 95% prediction intervals. The bias is given by

$$\text{Bias} = \frac{1}{m} \sum_{i=1}^{m} \Big(Z(\mathbf{s}_i^*) - Z(\mathbf{s}_i) \Big), \tag{4.7}$$

where a bias of zero indicates the predictions are the correct size on average. The RMSPE is given by

$$\text{RMSPE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \Big(Z(\mathbf{s}_i^*) - Z(\mathbf{s}_i) \Big)^2}, \tag{4.8}$$

and for unbiased predictions, it measures the amount of variation in the predictions around the true value, with smaller values indicating more precise estimation. Finally, a 95% prediction interval is computed for each predicted NO_2 concentration, and the coverage probability of a model is the percentage of these prediction intervals that contain the true value. The prediction intervals are the correct width if 95% of these intervals contain the true value.

4.3.3 Inference and McMC algorithm

The McMC simulation algorithm for model (4.1) produces a set of J samples for each of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \nu^2, \sigma^2, \rho)$, based on a mixture of Gibbs sampling and Metropolis-Hastings steps as discussed in Chapter 2 Section 2.3.2. The results are based on 10,000 posterior samples generated from one Markov chain, which has been burnt-in until convergence by assessing the stability of trace plots of the McMC samples for selected parameters (see Figure 4.5). The algorithm produces posterior distributions for each of the model parameters in $\boldsymbol{\theta}$, and the joint posterior distribution of $\boldsymbol{\theta}$ is given by:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\phi}, \nu^2, \sigma^2, \rho) \quad &\propto \quad p(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \nu^2)p(\boldsymbol{\beta})p(\boldsymbol{\phi}|\sigma^2, \rho)p(\sigma^2)p(\nu^2)p(\rho), \\ &= \quad \mathrm{N}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi}, \nu^2\sigma^2 I)\mathrm{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \mathbf{V}_\beta)\mathrm{N}(\boldsymbol{\phi}|\mathbf{0}, \sigma^2\boldsymbol{\Sigma}(\rho)) \\ &\quad \times \quad \Gamma^{-1}(\sigma^2|a, b)\mathrm{U}(\nu^2|0, 1)\mathrm{DU}(\rho_1, \ldots, \rho_r), \end{aligned} \tag{4.9}$$

where Γ^{-1} denotes the Inverse-Gamma distribution, U denotes the uniform distribution, and DU denotes the discrete uniform distribution. The full conditional distributions for each of the individual model parameters are described below.

$\boldsymbol{\beta}$ - regression parameters

The full conditional distribution for the regression parameters, $\boldsymbol{\beta}$, is a combination of the data likelihood given in equation (4.1) and the prior distribution for $\boldsymbol{\beta}$. Since this a combination of two Gaussian distributions, the resulting full conditional distribution for $\boldsymbol{\beta}$ is a Gaussian distribution, which is a known standard statistical distribution that can be simulated from utilising Gibbs sampling (see Chapter 2 Section 2.3.2 for further details on Gibbs sampling). The full conditional distribution is as follows:

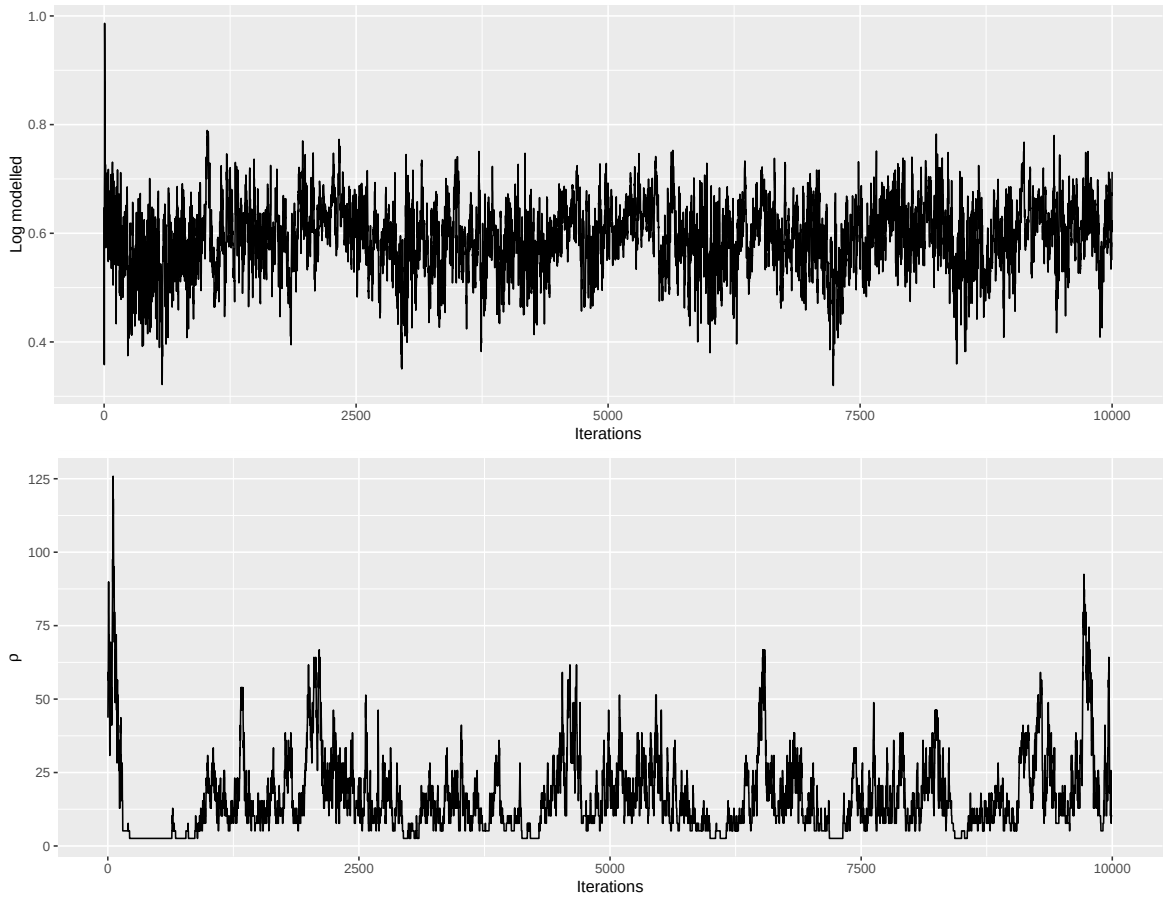

Figure 4.5: *Trace plots of selected model parameters under the full Bayesian model (model 1), where the top plot refers to the regression parameter for log modelled, and the bottom plot refers to the spatial decay parameter ρ. Similar results were found for all parameters across all models.*

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\phi},\nu^2,\sigma^2,\mathbf{Z}) \quad &= \quad \mathrm{N}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\phi},\nu^2\sigma^2\mathbf{I})\ \mathrm{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta,\mathbf{V}_\beta), \qquad (4.10)\\
&\propto \quad \exp\left(-\frac{1}{2\nu^2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right)\\
&\quad \times \quad \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^\top\mathbf{V}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\right),\\
&\propto \quad \exp\left\{-\frac{1}{2\nu^2\sigma^2}\left(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Z}-2\boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{\phi}\right)\right\}\\
&\quad \times \quad \exp\left(\boldsymbol{\beta}^\top\mathbf{V}_\beta^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta\right),\\
&\propto \quad \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top\left(\frac{\mathbf{X}^\top\mathbf{X}}{\nu^2\sigma^2}+\mathbf{V}_\beta^{-1}\right)\boldsymbol{\beta}\right]\right.\\
&\quad \left. -2\boldsymbol{\beta}^\top\left[\frac{\mathbf{X}^\top\mathbf{Z}}{\nu^2\sigma^2}+\frac{\mathbf{X}^\top\boldsymbol{\phi}}{\nu^2\sigma^2}+\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta\right]\right\},\\
&\sim \quad \mathrm{N}(\tilde{\boldsymbol{\beta}},\tilde{\Sigma}),
\end{aligned}
$$

where $\mathbf{I}$ is the identity matrix, $\tilde{\boldsymbol{\beta}} = \left[\frac{\mathbf{X}^\top\mathbf{X}}{\nu^2\sigma^2}+\mathbf{V}_\beta^{-1}\right]^{-1}\left[\frac{\mathbf{X}^\top(\mathbf{Z}-\boldsymbol{\phi})}{\nu^2\sigma^2}+\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta\right]$ and $\tilde{\Sigma} = \left[\frac{\mathbf{X}^\top\mathbf{X}}{\nu^2\sigma^2}+\mathbf{V}_\beta^{-1}\right]^{-1}$.

$\boldsymbol{\phi}$ - spatial random effects

Similarly to $\boldsymbol{\beta}$, the full conditional distribution for the spatial random effects, $\boldsymbol{\phi}$, is also a Gaussian distribution, since it is a combination of the data likelihood model in equation (4.1) and the multivariate Gaussian prior distribution for $\boldsymbol{\phi}$, which again is Gibbs sampled. The resulting full conditional distribution for $\boldsymbol{\phi}$ is as follows:

$$
\begin{aligned}
p(\boldsymbol{\phi}|\boldsymbol{\beta},\nu^2,\sigma^2,\rho,\mathbf{Z}) &= \mathrm{N}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\phi},\nu^2\sigma^2\mathbf{I})\mathrm{N}(\boldsymbol{\phi}|\mathbf{0},\sigma^2\boldsymbol{\Sigma}(\rho)), \qquad (4.11)\\
&\propto \exp\left(-\frac{1}{2\nu^2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right)\times\\
&\quad \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}\right),\\
&\propto \exp\left(-\frac{1}{2\nu^2\sigma^2}(\boldsymbol{\phi}^\top\mathbf{I}\boldsymbol{\phi}-2\boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{\phi}-2\mathbf{Z}^\top\boldsymbol{\phi})-\left(\frac{1}{2\sigma^2}\boldsymbol{\phi}^\top\mathbf{I}\boldsymbol{\phi}\right)\right),\\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\boldsymbol{\phi}^\top\mathbf{I}\boldsymbol{\phi}}{\nu^2\sigma^2}+\frac{\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}}{\sigma^2}-2\boldsymbol{\phi}^\top\left(\frac{\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}}{\nu^2\sigma^2}\right)\right]\right),\\
&\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\phi}^\top\left(\frac{1}{\nu^2\sigma^2}\mathbf{I}+\frac{\boldsymbol{\Sigma}(\rho)^{-1}}{\sigma^2}\right)\boldsymbol{\phi}-2\boldsymbol{\phi}^\top\frac{(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})}{\nu^2\sigma^2}\right]\right),\\
&\sim \mathrm{N}(\tilde{\boldsymbol{\phi}},\tilde{\Lambda}),
\end{aligned}
$$

where $\tilde{\boldsymbol{\phi}}=\sigma^2\left[\frac{1}{\nu^2}\mathbf{I}+\boldsymbol{\Sigma}(\rho)^{-1}\right]^{-1}\left[\frac{\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}}{\nu^2\sigma^2}\right]$ and $\tilde{\Lambda}=\sigma^2\left[\frac{1}{\nu^2}\mathbf{I}+\boldsymbol{\Sigma}(\rho)^{-1}\right]^{-1}$.

σ^2 - spatial variance

The spatial variance parameter for the random effects is given by σ^2 and is present in three models: the data likelihood model given by equation (4.1), the multivariate Gaussian distribution for the random effects, $\boldsymbol{\phi}$, given by (4.2), and its own Inverse-Gamma prior distribution. Therefore, the resulting full conditional distribution for σ^2 is a combination of these three models, which produces an Inverse-Gamma distribution that can be Gibbs sampled. It is given by:

$$
\begin{aligned}
p(\sigma^2|\boldsymbol{\beta},\boldsymbol{\phi},\nu^2,\rho,\mathbf{Z}) &= \mathrm{N}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\phi},\nu^2\sigma^2\mathbf{I})\ \mathrm{N}(\boldsymbol{\phi}|\mathbf{0},\sigma^2\boldsymbol{\Sigma}(\rho))\ \Gamma^{-1}(\sigma^2|a,b), \qquad (4.12)\\
&\propto (\sigma^2)^{-\frac{m}{2}}(\sigma^2)^{-\frac{m}{2}}(\sigma^2)^{-(a+1)}\times\\
&\quad \exp\left(-\frac{1}{2\nu^2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right)\\
&\quad \times\ \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}\right)\exp\left(-\frac{b}{\sigma^2}\right),\\
&\propto (\sigma^2)^{-(m+a+1)}\times\\
&\quad \exp\left(-\frac{b+\frac{1}{2\nu^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})+\frac{1}{2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}}{\sigma^2}\right),\\
&\sim \Gamma^{-1}\left(m+a,\ b+\frac{1}{2\nu^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right.\\
&\qquad \left.+\frac{1}{2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}\right).
\end{aligned}
$$

ν^2 - variance parameter

The level of the automatic monitor and diffusion tube measurement error (nugget) is specified by $\nu^2\sigma^2$, where the full conditional distribution for the variance parameter ν^2 is as follows:

$$\begin{aligned} p(\nu^2|\boldsymbol{\beta},\boldsymbol{\phi},\sigma^2,\mathbf{Z}) &= \mathrm{N}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\phi},\nu^2\sigma^2\mathbf{I})\ \mathrm{Uniform}(\nu^2|0,1), \\ &\propto (\nu^2)^{-\frac{m}{2}}\exp\left(-\frac{1}{2\nu^2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right) \\ &\propto (\nu^2)^{-(\frac{m}{2}-1+1)}\exp\left(-\frac{1}{2\nu^2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right), \\ &\sim \Gamma^{-1}\left(\frac{m}{2}-1,\frac{1}{2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})^\top(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right). \end{aligned} \tag{4.13}$$

This produces an Inverse-Gamma distribution for ν^2, which is sampled using Gibbs sampling.

ρ - spatial decay parameter

The full conditional distribution for the spatial decay parameter, ρ, is a combination of the multivariate Gaussian distribution for the spatial random effects and the discrete Uniform prior specified for ρ. The resulting full conditional distribution is not a known standard statistical distribution and is therefore sampled using the Metropolis-Hastings algorithm discussed in Chapter 2 Section 2.3.2. The full conditional distribution for ρ is given by

$$\begin{aligned} p(\rho|\boldsymbol{\phi},\sigma^2) &= \mathrm{N}(\boldsymbol{\phi}|\mathbf{0},\sigma^2\boldsymbol{\Sigma}(\rho))\ \mathrm{DU}(\rho_1,\ldots,\rho_r), \\ &\propto |\boldsymbol{\Sigma}(\rho)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}\right)\ I[\rho\in\{\rho_1,\ldots,\rho_r\}], \\ \ln[p(\rho|-)] &\propto -\frac{1}{2}\ln|\boldsymbol{\Sigma}(\rho)| - \frac{1}{2\sigma^2}\boldsymbol{\phi}^\top\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{\phi}. \end{aligned} \tag{4.14}$$

As stated in Section 4.3.1, a discrete uniform prior is specified for ρ to ensure that the spatial correlation matrix, $\boldsymbol{\Sigma}(\rho)$, need only be inverted a specified number of times rather than at every J iteration of the McMC algorithm, which would dramatically increase the computational time of the algorithm. Therefore, based on the recommendation from Diggle & Ribeiro (2007) values for ρ are generated within the range $(0.01, 2\max(\mathbf{D}))$, where $\max(\mathbf{D})$ relates to the maximum distance between any two locations. The proposal, ρ^* is then sampled from within this range, and the acceptance probability of a move from $\rho^{(i)}$ to ρ^* is given by $\min\left(1, \frac{p(\rho^*)}{\rho^{(i)}}\right)$.

4.3.3.1 Algorithm

The model parameters are updated using a combination of a Metropolis-Hastings step for ρ and Gibbs sampling for the remaining parameters. The spatial prediction algorithm proposed here is as follows:

1. Calculate Euclidean Distance matrices for the spatial correlation matrix for the observed locations $\boldsymbol{\Sigma}(\rho)$, the prediction locations $\boldsymbol{\Sigma}^*(\rho)$, and between both observed and prediction locations $\mathbf{C}_Z(\mathbf{s}^*, \rho)$.

2. Create the vector of $r = 50$ fixed values for the spatial decay parameter, ρ, based on $(0.01, 2\max(\mathbf{D}))$.

3. Calculate the corresponding spatial correlation matrices (inverse and determinant) from the distance matrices created in Step 1 and the proposed candidate values for ρ in Step 2.

4. Compute starting values for all model parameters based on their prior distributions, where a starting value for ρ is sampled randomly from its predetermined set of values.

5. Perform the Gibbs Sampling and Metropolis-Hastings step for the model parameters and the spatial prediction for $j = 1, \ldots, J$ samples. The model parameters are sampled as follows:

   (a) Sample the block of regression parameters, $\boldsymbol{\beta}^{(j)}$, from its full conditional distribution given by $p(\boldsymbol{\beta}|\boldsymbol{\phi}^{(j-1)}, \nu^{2(j-1)}, \sigma^{2(j-1)}, \mathbf{Z})$.

   (b) Sample the spatial variance parameter, $\sigma^{2(j)}$, from its full conditional distribution given by $p(\sigma^2|\boldsymbol{\beta}^{(j)}, \boldsymbol{\phi}^{(j-1)}, \nu^{2(j-1)}, \rho^{(j-1)}, \mathbf{Z})$.

   (c) Sample the spatial decay parameter, $\rho^{(j)}$, from its full conditional distribution $p(\rho|\boldsymbol{\phi}^{(j-1)}, \sigma^{2(j)})$.

   (d) Sample the variance parameter, $\nu^{2(j)}$, from its full conditional distribution given by $p(\nu^2|\boldsymbol{\beta}^{(j)}, \boldsymbol{\phi}^{(j-1)}, \sigma^{2(j)}, \mathbf{Z})$.

   (e) Sample the random effects, $\boldsymbol{\phi}^{(j)}$, from its full conditional distribution given by $p(\boldsymbol{\phi}|\boldsymbol{\beta}^{(j)}, \nu^{2(j)}, \sigma^{2(j)}, \rho^{(j)}, \mathbf{Z})$.

   (f) Sample the random effects, $\boldsymbol{\phi}^{*(j)}$, at the N prediction locations $\mathbf{s}^* = (\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*)$ from its full conditional distribution $p(\boldsymbol{\phi}^*|\boldsymbol{\beta}^{(j)}, \nu^{2(j)}, \sigma^{2(j)}, \rho^{(j)}, \mathbf{Z})$.

   (g) Generate the predicted value of $Z(\mathbf{s}^*)^{(j)}$ for the N prediction locations given by equation (4.6).

4.4 Results

This section presents results from applying the statistical fusion model proposed in Section 4.3 to the Glasgow case study outlined in Section 4.2. Section 4.4.1 presents a validation study comparing the appropriateness of the proposed model against a number of alternative models in terms of both model structure and covariate choice. Section 4.4.2 demonstrates the advantages of using the diffusion tube data for fine scale spatial prediction by comparing predictive accuracy against using the automatic monitors alone. Finally, Section 4.4.3 uses the best performing model from the previous two sections to predict yearly average NO_2 concentrations at a 1 kilometre grid square resolution across West Central Scotland, with associated 95% prediction intervals.

4.4.1 Validation study 1: model structure and covariate choice

In this validation study, the predictive performances of a number of different model specifications are compared, focusing on the utility of allowing for spatial autocorrelation in the data, the approach to parameter estimation adopted for the model, and the choice of covariates.

The results of the validation study are presented in Table 4.3, for nine different models and are compared in terms of bias, RMSPE and coverage probability (as detailed in Section 4.3.2). The top panel of the table compares the utility of allowing for spatial autocorrelation in the data and the estimation approach taken, while the bottom panel shows a sensitivity analysis to the choice of covariates. In all cases the models are unbiased, as the biases are all close to zero, ranging between -0.0001 and 0.356. Model 1 is the full Bayesian model described in Section 4.3, which includes the log-transformed modelled NO_2 concentrations (*log modelled*), an indicator for the type of measured data (automatic monitor or non-automatic diffusion tube, *monitor/tube*), and the local environment in which each observation resides (e.g., roadside, urban background, etc, *environment*) as covariates. Model 2 is the same as Model 1 except that inference is performed using restricted maximum likelihood estimation instead of Bayesian methods, and the RMSPE values are almost identical. The differences are in the coverage probabilities, with the Bayesian estimation having wider and more appropriate prediction intervals (coverages differ by around 1%) than under likelihood based estimation. This small difference occurs as, when using restricted maximum likelihood, the estimated model parameters are assumed to be fixed and known when making the predictions, thus underestimating the amount of uncertainty in the data. In contrast, the Bayesian model allows for uncertainty in the estimated model parameters when making predictions, thus explaining its wider prediction intervals. Model 3 also uses maximum likelihood estimation, but naively ignores the spatial autocorrelation present in the data. This model shows around a 5% increase in RMSPE compared with Model 1, suggesting that ignoring the spatial autocorrelation in the data results in poorer

predictive performance.

Table 4.3: *Bias (μgm^{-3}), RMSPE (μgm^{-3}) and coverage probability (%) results all models compared in this section. The top panel displays the models with different estimation methods (Models 1-3), while the bottom panel displays the results for the Bayesian models containing differing covariate combinations (Models 4-9).*

Model	Bias	RMSPE	Coverage
1	0.010	0.257	93.089
2	0.005	0.255	91.870
3	-0.0001	0.271	93.902
4	0.356	0.545	95.122
5	0.020	0.303	95.122
6	0.011	0.258	93.496
7	0.018	0.276	94.715
8	0.009	0.255	94.715
9	0.013	0.255	94.715

The bottom panel of Table 4.3 shows a comparison of different combinations of covariates, which are summarised below.

Model 1 - *log modelled + monitor/tube + environment*

Model 4 - *log modelled*

Model 5 - *log modelled + monitor/tube*

Model 6 - *log modelled + environment*

Model 7 - *monitor/tube + environment*

Model 8 - *log modelled + monitor/tube + environment + easting + northing + log modelled:easting + log modelled:northing*

Model 9 - *log modelled + environment + easting + northing + log modelled:easting + log modelled:northing*

In all cases, the Bayesian fusion model described in Section 4.3 is used. These results show two main points. Firstly, the *log modelled* and *environment* variables are important for accurate NO_2 prediction, which is evidenced by an increase in RMSPE for Models 4, 5 and 7 compared with Model 1. The bias and RMSPE is much greater for Model 4 compared to Model 1 and Models 5-7. This is because Model 4 did not include any covariates to distinguish between measurements made in different environments, i.e., roadside, kerbside, rural or urban background locations. The *log modelled* variable is a spatially smooth covariate with no adjustment for the environment, so

it cannot capture higher measurements at the roadside and lower background measurements, and therefore, tends to overestimate the pollutant concentrations, which is evidenced by its higher bias and RMSPE. The importance of the *log modelled* covariate is clear, while the *environment* variable is important because it distinguishes between observations at roadside and background environments, which will have a large impact on the measured NO_2 value since it is largely driven by traffic sources. Secondly, including the *monitor/tube* variable does not lead to improved NO_2 prediction, as the RMSPE of Model 6 is 0.258 compared to 0.257 for Model 1. This can also be shown in Table 4.4, which displays the posterior medians and 95% credible intervals for each of the covariates for Model 1. NO_2 concentrations recorded by automatic monitors are slightly higher compared to NO_2 concentrations recorded by diffusion tubes as the posterior median is positive. However, the relationship is very weak as the 95% CI's lower bound is close to zero. Furthermore, rural, special and urban background sites have substantially lower NO_2 concentrations compared to kerbside sites; however, even though roadside sites have lower NO_2 pollution levels compared to kerbside sites, the relationship is quite weak as the upper bound for the credible interval is just below one, which is not surprising since roadside and kerbside sites both measure pollution at the roadside.

The bias and RMSPE in the modelled concentrations are also computed, thus allowing the improvement in predictive performance from the above models to be observed. Since the modelled concentrations are ambient, background concentrations, an adjustment for sites measured at roadside and kerbside environments was included. The results show that even after adjusting the modelled concentrations for roadside and kerbside environments (*roadside/otherwise*), the modelled concentrations are not as good for predicting NO_2 concentrations compared to the best Model with a RMSPE of 0.337 (and a bias of -0.068) compared to 0.258 for Model 6.

Each of the models described above assume the effect of the modelled concentrations is constant across space. This necessarily might not be the case as the effect may vary depending on the spatial location. Therefore, to allow flexibility in the effect of the modelled concentrations to vary across space, Models 8 and 9 contain an interaction term between the *log modelled* variable and the easting and northing coordinates of the location of the automatic monitors and diffusion tubes, given as *log modelled:easting* and *log modelled:northing*. Model 9 is the same as Model 8 except it does not contain the *monitor/tube* covariate. Both models are unbiased, and have the same RMSPE of 0.255 and coverage probability of 94.715%, indicating again that the *monitor/tube* variable does not lead to improved NO_2 prediction. Even though the bias in Model 9 (0.013) is slightly higher compared to Model 8 (0.009), Model 9 is a better model compared to Model 6 as it does not treat the effect of the modelled concentrations to be constant across space. Furthermore, the RMSPE decreases by around 1% and

Table 4.4: *Posterior medians and 95% credible intervals (CI) for selected parameters of Model 1, which is the full Bayesian model with log modelled, monitor/tube and environment as covariates. The diffusion tubes were taken as the reference category for monitor/tube and kerbside was taken as the reference category for environment. Results are also shown for the spatial variance σ^2, noise-to-signal ratio ν^2 and spatial decay parameter ρ.*

Variable	Posterior median	95% CI
Intercept	1.900	(1.564, 2.259)
Log modelled	0.594	(0.459, 0.708)
Monitor	0.125	(0.027, 0.238)
Roadside	-0.150	(-0.258, -0.042)
Rural	-1.021	(-1.585, -0.488)
Special	-0.390	(-0.630, -0.147)
Urban background	-0.531	(-0.659, -0.407)
σ^2	0.057	(0.024, 0.077)
ν^2	0.232	(0.064, 1.819)
ρ	12.852	(2.578, 53.946)

the coverage probability is closer to the nominal value of 95%. Therefore, the final model considered here is Model 9, as it is more flexible compared to Model 6, and has improved performance, mainly in terms of coverage probability. The effect of the modelled concentrations was also considered to vary as a quadratic surface in location by adding into Model 9 the covariates *log modelled:easting*^{2} + *log modelled:northing*^{2}; however, this did not improve spatial NO_2 prediction as the RMSPE (0.258) was higher compared to Model 9 (0.255).

4.4.2 Validation study 2: data source

The second validation study investigates the effectiveness of using the diffusion tube data in addition to the automatic monitoring data for fine scale spatial prediction. Model 9 is used throughout this section, as Section 4.4.1 showed it had the best overall performance. In common with Section 4.4.1, leave-one-out cross-validation is used to assess predictive accuracy, again using bias, RMSPE and coverage probabilities to quantify prediction performance. Model 9 is fit to two subsets of the data: one where only the 16 automatic monitors are used as the response, and one where only the 230 diffusion tubes are used as the response. In each case Model 9 is used to predict each of the 246 observations in turn.

The results of the second validation study are displayed in Table 4.5 for the two different subsets of the measured pollutant data, along with the full data set containing all 246 observations from both the automatic monitors and diffusion tubes. In common with the results from the first validation study in Section 4.4.1, all three sets of data

are unbiased as the biases are all close to zero, ranging from 0.009 to 0.266. However, the predictive performance from using only the automatic monitors is markedly poorer than using either just the diffusion tubes or all the observations. This is evidenced by both its RMSPE and coverage probability. The RMSPE from using the automatic monitors only is 0.478, which is 48% (RMSPE of 0.249) and 47% (RMSPE of 0.255) greater than the corresponding values from using the diffusion tubes only and the combined data set respectively. Additionally, the coverage probability when using the automatic monitors alone is over 99.5%, which is larger than the nominal 95% levels. This high coverage probability suggests that the prediction intervals are too wide, most likely due to a lack of data provided by the automatic monitors, of which there are only 16, thus resulting in poorer parameter estimation and higher uncertainty.

Table 4.5: *Bias (μgm^{-3}), RMSPE (μgm^{-3}) and coverage probabilities (%) for the leave-one-out cross-validation of applying Model 9 to the three different sources of data. One data set containing only the automatic monitors, one data set containing only the diffusion tubes, and one data set containing a combination of both automatic monitors and diffusion tubes.*

Data source	Bias	RMSPE	Coverage
Monitors	0.266	0.478	99.594
Tubes	0.009	0.249	95.122
Monitors & Tubes	0.013	0.255	94.715

In contrast, the coverage probabilities from using just the diffusion tubes and all the measured data are close to their nominal 95% levels, while the RMSPE for the diffusion tube only model is 0.249 and 0.255 for the combined model. These results suggest that using the automatic monitors in addition to the diffusion tubes does not lead to better predictive performance compared to using the diffusion tubes alone, as the two sets of results are essentially the same after allowing for random error. The reason for this is that some of the automatic monitors are co-located with the diffusion tubes so when the automatic monitors are included with the diffusion tubes, there is not a large increase in the number of observed data points. These results, therefore, demonstrate the effectiveness of using the diffusion tube data for predicting the NO_2 concentrations at a fine spatial scale.

4.4.3 NO_2 prediction

The model chosen to predict NO_2 concentrations at each 1km $\times$1km grid box was Model 9: the Bayesian fusion model including both automatic monitors and diffusion tubes, with the log-transformed modelled concentrations and the local environment in which each automatic monitor and diffusion tube resides as covariates. The effect of the modelled concentrations was also allowed to vary across space by including an interaction term between the modelled concentrations and the easting and northing

coordinates of the monitors and tubes. For spatial prediction purposes, grid boxes were predicted as urban background or rural. The remainder of the local environment locations (kerbside, roadside, and special) were not considered here because the NO_2 concentrations to be predicted are averages over a 1km $\times$ 1km grid box, meaning that NO_2 measurements produced from the roadside are averaged out across the grid box. Furthermore, when estimating the effect of air pollutants of ill health, roadside concentrations are not representative of exposure levels as people tend not to spend large proportions of their time next to a road. The urban-rural variable discussed in Section 4.2.3 is used instead to predict each grid box as urban background or rural.

Figure 4.6 displays the final predicted NO_2 concentrations across West Central Scotland and their associated standard errors. The median of the 10,000 posterior samples was taken to be the Bayesian point estimate for these predicted NO_2 concentrations. In common with the modelled concentrations in Figure 4.3, the City of Glasgow and the main road network are easily distinguishable. Summary statistics for the predicted NO_2 concentrations, their standard errors and the modelled concentrations are shown in Table 4.6 separately for urban and rural areas. These statistics highlight that the median concentration predicted from Model 9 is 23.000 μgm^{-3} for urban areas and 12.060 μgm^{-3} for rural areas, while for the modelled concentrations the median value is 11.680 and 4.849 μgm^{-3} for urban and rural areas respectively. Therefore as shown in Table 4.3 that the predictions from Model 9 are unbiased, the modelled concentrations are likely to be underestimating the level of NO_2 across West Central Scotland. However, the spatial pattern in the two sets of data is similar, with a Pearson's correlation coefficient of 0.923 between the predictions from Model 9 and the modelled concentrations. Figure 4.7 highlights the strong agreement between the predicted NO_2 concentrations and the modelled NO_2 concentrations.

The next chapter in this thesis presents a new study of NO_2 concentrations that are combined with cardio-respiratory mortality data in order to estimate the pollutant-health relationship in West Central Scotland. This requires predicted NO_2 concentrations for years 2007 to 2012, which are now described in this section. The number of automatic monitors and diffusion tubes were not fixed over the seven year period between 2006 and 2012. Table 4.7 displays the numbers of automatic monitors and diffusion tubes in each year of the study period. The number of automatic monitors increased over the seven years from 16 in 2006 to 24 in 2012, while the numbers of diffusion tubes increased from 230 in 2006 to 299 in 2012.

However, one has to bear in mind that the numbers of automatic monitors and diffusion tubes may not mirror the same locations across the years since automatic monitors and diffusion tubes become faulty, which can result in the closure and introduction of new sites across the study region. The spatial distribution of the automatic

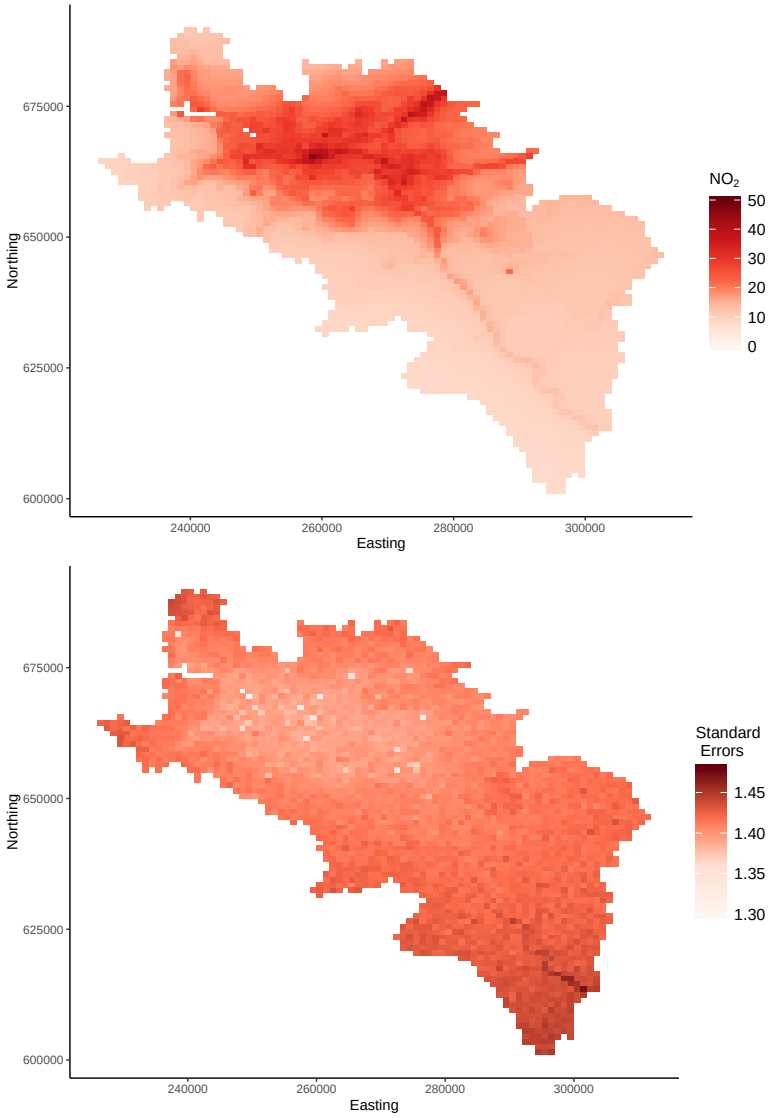

Figure 4.6: *The top map shows the 2006 predicted NO_2 (μgm^{-3}) concentrations from Model 9 across West Central Scotland, while the bottom map shows the corresponding standard errors.*

monitors and diffusion tubes for the remaining years is similar to that seen in Figure 4.2 for the year 2006.

Model 9 was used to predict the NO_2 concentrations separately for the remaining years using the methodology described above, and summary statistics are presented in Table 4.8. The ranges of predicted NO_2 concentrations are similar across the years, with 2010 having the highest maximum concentration of 56.980 μgm^{-3}. Furthermore, spatial maps of the predicted NO_2 concentrations and their corresponding standing errors are presented in Appendix A, and all years exhibit similar spatial patterns. In contrast, the year 2012 has higher levels of predicted NO_2 concentrations across West Central Scotland compared to previous years.

Table 4.6: *Summary statistics for the 2006 modelled and predicted NO_2 (μgm^{-3}) concentrations from Model 9 with associated standard errors separately for urban and rural areas.*

	Modelled NO_2	Predicted NO_2	Standard Errors
Urban areas			
Min	3.207	12.570	1.314
25th Percentile	7.985	18.300	1.394
Median	11.680	22.650	1.399
Mean	12.040	23.000	1.400
75th Percentile	15.230	27.42	1.406
Max	34.760	46.400	1.439
Rural areas			
Min	3.021	8.028	1.379
25th Percentile	4.268	10.230	1.411
Median	4.849	12.060	1.417
Mean	5.575	13.020	1.418
75th Percentile	6.207	14.060	1.424
Max	18.090	32.090	1.472

4.5 Discussion

This chapter demonstrates that improvements in the accuracy of fine scale spatial prediction of NO_2 concentrations can be made by using diffusion tube data in addition to the commonly-used automatic monitors. Diffusion tubes are relatively inexpensive and thus, more prevalent than automatic monitors in many urban environments, and the subsequent large increase in the number of spatial locations at which NO_2 is measured leads to improvements in predictive performance. The Bayesian geostatistical fusion model proposed links the measured and modelled NO_2 concentrations via a regression relationship, and is similar to existing downscaling models used in the literature (Berrocal et al., 2010a,b). The model performs fine scale spatial NO_2 predictions that are unbiased and have appropriate width prediction intervals. Thus, this modelling framework should be useful for predicting NO_2 concentrations in other urban environments.

The results from this chapter have illustrated three key points. Firstly, using the diffusion tube data in addition to the automatic monitoring data enhances the predictive performance of fine scale NO_2 concentrations, compared to using the automatic monitors alone. This is evidenced by a 47% reduction in RMSPE when utilising both sources of NO_2 concentrations. This reduction in RMSPE is due to the increase in the number of observations used, resulting in more accurate parameter estimation and lower uncertainty. Furthermore, the bias reduced by a factor of 10 and the coverage

Table 4.7: *Total number of automatic monitors and diffusion tubes for the years 2006 to 2012.*

Year	Automatic monitors	Diffusion tubes
2006	16	230
2007	18	257
2008	23	252
2009	23	305
2010	25	290
2011	23	311
2012	24	299

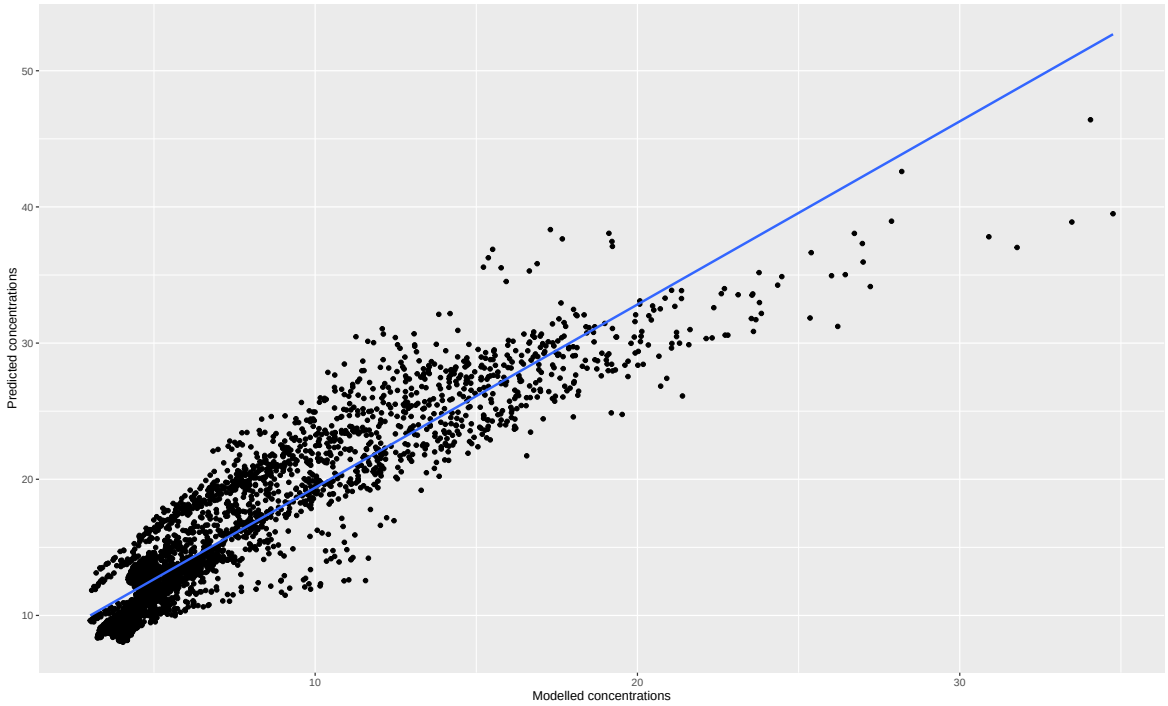

Figure 4.7: *Scatter plot highlighting the high agreement between the predicted NO_2 concentrations and the modelled NO_2 concentrations.*

probability improved by 5% when using both sets of measured pollutant data. The latter is important because using monitoring data alone resulted in predictive uncertainty that was too high. This was due to the considerably fewer automatic monitors compared to the diffusion tubes (16 compared to 230). Secondly, using the modelled concentrations leads to improved spatial prediction, as a model containing the modelled concentrations surpassed the model without the modelled concentrations, with RMSPEs of 0.257 and 0.276 respectively, which is an increase of 7% in predictive accuracy. Finally, it is important to allow for spatial autocorrelation in the data, as the RMSPE increased by 5% compared to the model that did not take into account spatial autocorrelation. Furthermore, Bayesian methods allow for better uncertainty quantification than likelihood based estimation, as the coverage probability is closer to the nominal 95% level. The chosen model (Model 9) allowed the effect of the mod-

Table 4.8: *Summary statistics for Model 9 predicted NO_2 concentrations for the years 2007 to 2012.*

Year	Min	25 Percentile	Median	Mean	75th Percentile	Max
2007	7.002	9.657	11.980	14.910	18.970	50.170
2008	9.075	11.870	14.120	17.020	20.990	49.810
2009	6.487	9.648	12.330	15.720	21.030	51.980
2010	12.840	16.000	19.130	21.020	25.300	56.430
2011	8.325	11.460	14.440	16.730	21.300	46.720
2012	15.420	20.290	22.290	23.250	25.530	43.260

elled concentrations to vary across space, which showed a slight improvement over the model that assumed the effect was constant (Model 6: RMSPE of 0.255 compared to 0.258). The main difference between these two models is in the coverage probability, which was closer to the 95% nominal level for the model that considered the effect to vary across space. In absolute terms, the results do not show large differences between a model without using the diffusion tubes and a model with the diffusion tubes, but a 47% increase in performance highlights the utility of this approach. Furthermore, the model that only makes use of the automatic monitors for prediction, had credible intervals that were far too wide, with a coverage probability of 99.594%. This was due to the small number of automatic monitors and hence this model was poorer than the model using both automatic monitors and diffusion tubes.

The methodology proposed here has a number of limitations. The temporal resolution for the study was yearly, but it would be more desirable to be able to apply the same methodology to higher resolution time periods, such as daily. However, the diffusion tube data are only available as monthly averages, preventing the use of this approach at finer temporal scales. Furthermore, background NO_2 concentrations were predicted using the modelled NO_2 data at a 1km grid square resolution, and thus the predictions are background concentrations that do not include local sources, such as roads. In addition, NO_2 concentrations cannot be predicted at a finer spatial scale as the modelled concentrations are only available at the 1km grid square resolution. Irrespective of these limitations the predicted concentrations are utilised in the following two chapters to investigate the relationship between NO_2 and ill health in West Central Scotland.

Chapter 5

How robust are the estimated effects of air pollution on health? Accounting for model uncertainty using Bayesian model averaging

5.1 Introduction

As discussed in Chapter 3, the health impacts of exposure to both long-term (chronic) and short-term (acute) air pollution have been much researched, where the long-term health impact of air pollution is most often estimated from cohort studies (see Cesaroni et al., 2014). However, cohort studies are expensive and time consuming to implement, and it may take years before results are available. This has led to spatial ecological study designs being implemented instead (see Lee et al., 2009; Maheswaran et al., 2005a), where routinely available small area data can be used, which makes the implementation of such studies much quicker. However, because they are conducted on aggregate data rather than at the individual level, they cannot be used as a means of determining causation between exposure to air pollution and subsequent ill health. Further details of this study design is discussed in Chapter 3 Section 3.3

As with all statistical modelling endeavours, estimating the effects of air pollution on ill health requires a number of modelling choices to be made, which are likely to affect the results. This variation in effect estimates due to model uncertainty is typically ignored, and results from a single 'final' model are often presented. However, it is likely to be crucial in this context, because the estimated effect sizes are small and their significance will depend on the final model chosen (as highlighted in Table 5.3), thus it is likely that statistically significant or non-significant results could be presented depending on the choices made by the investigators.

In this chapter, the impact of three such modelling choices is investigated, namely estimation of NO_2 concentrations, the measure of socio-economic deprivation used, and the method for controlling residual spatial autocorrelation. This chapter utilises the fine scale NO_2 concentrations produced from the geostatistical fusion model developed in Chapter 4, while also comparing its health effects to when the modelled concentrations are used instead, since this is what the majority of most spatial ecological studies use (Haining et al., 2010; Lee et al., 2009). Socio-economic deprivation is an important confounder in these studies, and existing studies have attempted to control for it using either individual-level measures, such as job seekers allowance or house price (Lee et al., 2014), or composite indexes, such as the Townsend index (Maheswaran et al., 2005a). This study makes use of the SIMD and its individual domains to account for deprivation, while keeping in line with previous studies. Finally, fitting a simple Poisson log-linear model to the data ignores any residual spatial autocorrelation. A common adjustment is to add a set of random effects represented by a globally smooth conditional autoregressive (CAR, Besag et al., 1991) prior to the linear predictor. However, Clayton et al. (1993); Hodges & Reich (2010); Reich et al. (2006), and Paciorek (2010) have shown this may lead to collinearity between the fixed and random effects, which can lead to poor estimation of the fixed effects. Thus, a number of extensions have been proposed to mitigate spatial confounding, such as the orthogonal smoothing approach by Hughes & Haran (2013), and the localised smoothing approach by Lee & Sarran (2015). This study utilises the approach by Hughes & Haran (2013), rather than Lee & Sarran (2015), since their method was strictly developed to mitigate against against spatial confounding.

This chapter will present a new study of NO_2 concentrations and cardio-respiratory mortalities in West Central Scotland, in which the robustness of the estimated pollutant-health effect sizes to these factors are quantified. A Bayesian model averaging (BMA, Hoeting et al., 1999; Raftery, 1995) approach to estimating the overall effect size, whilst accounting for model uncertainty is considered. This chapter is organised as follows. Section 5.2 describes the motivating study, along with descriptions of the disease, air pollutant and deprivation data. Section 5.3 presents the statistical models described above for taking into account residual spatial autocorrelation and estimating NO_2 concentrations. This section also presents the BMA methodology for combining the estimated air pollutant effects from the range of models considered. The results from the individual models and BMA are presented in Section 5.4, while Section 5.5 provides a concluding discussion.

5.2 Motivating study

The methodology developed in this chapter is motivated by a new epidemiological study investigating the health impact of long-term exposure to air pollution in West Central

Scotland, for the seven year period 2006 to 2012. This is the same study region as the fusion model study presented in Chapter 4 Section 4.2, but for the purposes of this study, West Central Scotland is partitioned into $m = 2089$ non-overlapping data zones comprising between 500 and 1000 (mean population = 800) residents of similar social characteristics. Data zones are the key small area geography in Scotland that are used for communicating and monitoring government statistics and to aid policy. Typical information relates to benefits, education, health, and area-level deprivation. These data zones cover the whole of Scotland (total of 6505) and individually nest within local authorities, where care has been taken to ensure they respect physical boundaries and are of a compact shape, albeit irregular. These data zones are described further at the Scottish Neighbourhood statistics website (SNS, http://www.sns.gov.uk/). The layout of the study region is presented in Figure 5.1, where the city of Glasgow is the set of small data zones in the middle north of the figure.

5.2.1 Disease data

The disease data comprise counts of the numbers of cardio-respiratory deaths (International Classification of Diseases, 10th Revision (ICD-10): I00-I99, J00-J99) within each of the 2089 data zones during the seven year period 2006 to 2012. These death records were obtained from National Records Scotland (https://www.nrscotland.gov.uk/) and held at the MRC/CSO Social and Public Health Sciences Unit (http://www.sphsu.mrc.ac.uk/). In order to take into account the heterogeneity of the population within each data zone in terms of their size and demographic structure, the expected numbers of cardio-respiratory deaths were calculated by indirect standardisation, using age- and sex-specific cardio-respiratory mortality rates for the whole of West Central Scotland as described in Chapter 2 Section 2.5.2. However, due to the low numbers of cardio-respiratory mortalities occurring in a single year (mean of 4.093 for 2006 as shown in Table 5.1), the cardio-respiratory deaths have been aggregated across the seven year period in order to increase the variation of deaths across the data zones. Furthermore, Table 5.1 shows that the distribution and total numbers of cardio-respiratory deaths do not change considerably over the seven year period, highlighting a lack of temporal variation. The spatial distribution of disease risk is shown in the bottom left panel of Figure 5.1, which displays the standardised mortality ratios (SMR, observed numbers/expected numbers) across West Central Scotland for the aggregated years 2006 to 2012. An SMR of 1.2 corresponds to a 20% increase in the risk of disease compared to what is expected. The highest SMRs are found in areas with the highest level of deprivation, and range between 0 and 4.747, with a mean SMR of 1.066, and a standard deviation of 0.440. Therefore, on average, there is a 6.6% increased risk of cardio-respiratory mortality relative to what is expected. Data zones have an SMR of zero when there have been no deaths. This mostly occurs in the centre of Glasgow, which consists mainly of shopping districts.

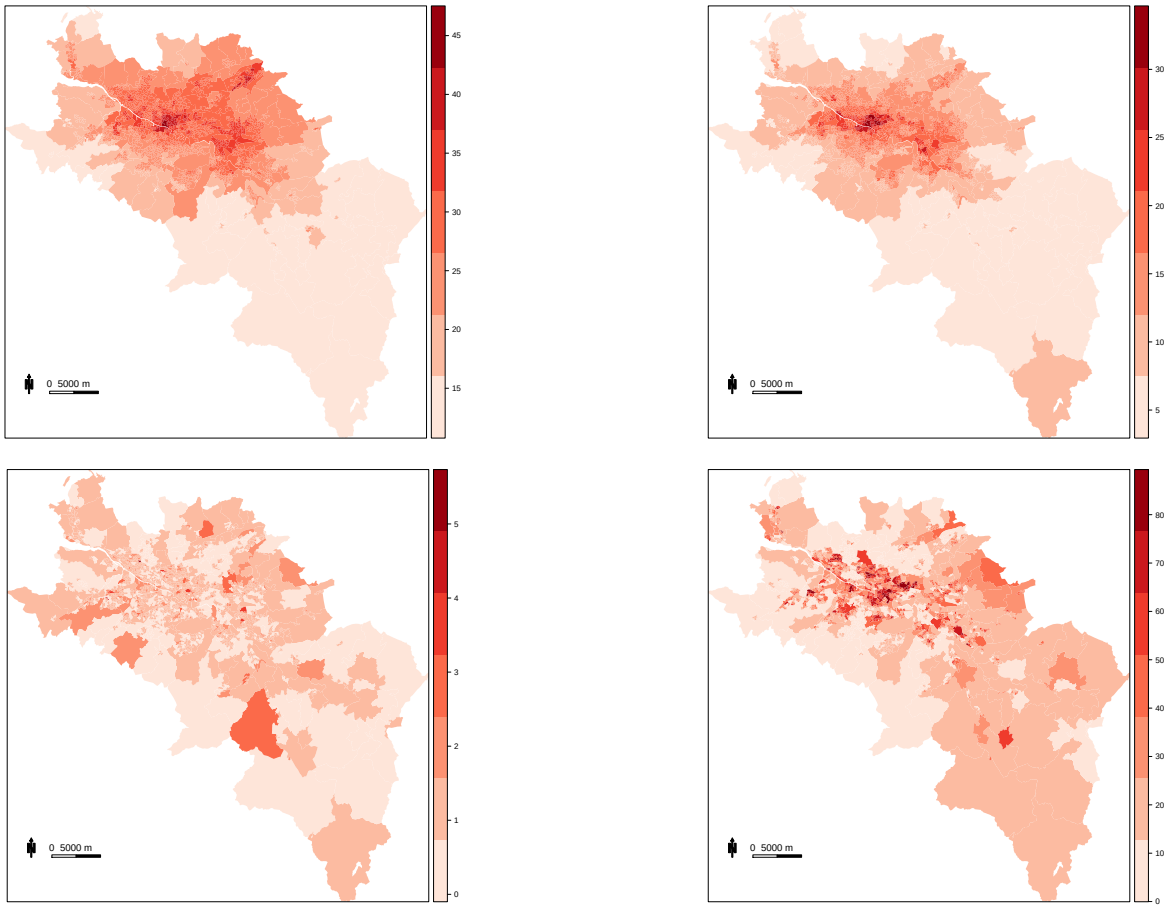

Figure 5.1: *Display of the data. The top left panel shows estimates from the statistical fusion model, while the top right panel shows background NO_2 concentrations provided by DEFRA both averaged across the seven year period 2006-2012. The bottom left panel displays the Standardised mortality ratio (SMR) for cardio-respiratory disease aggregated over 2006-2012, while the bottom right panel displays the SIMD score (without health domain), where a high score indicates deprivation and a low score indicates affluence.*

Table 5.1: *Summary statistics and total number of cardio-respiratory deaths separately for each year, and for aggregated years 2006-12 and across all data zones.*

Year	Min	25%	Median	Mean	75%	Max	Total
2006	0	2	3	4.093	6	46	8551
2007	0	2	3	4.095	5	39	8334
2008	0	2	3	3.989	5	39	8334
2009	0	1	3	3.714	5	36	7759
2010	0	1	3	3.649	5	31	7623
2011	0	1	3	3.569	5	40	7455
2012	0	1	3	3.623	5	38	7569
2006-2012	0	14	22	26.730	33	260	55,846

5.2.2 Air pollutant data

Both the modelled NO_2 concentrations and the NO_2 concentrations developed in this thesis are used to investigate its association with cardio-respiratory mortality. The first set of NO_2 concentrations are the modelled concentrations available as annual mean background concentrations at a 1km $\times$ 1km grid square resolution for the seven year period. These concentrations are temporally aggregated over the seven year period by averaging, then spatially aggregated to the data zone level using the following formula

$$\mathrm{NO}_{2_i} = \frac{\sum_{j=1}^{n_i} \exp(-d_{ij})\tilde{\mathrm{NO}}_{2j}}{\sum_{j=1}^{n_i} \exp(-d_{ij})}, \tag{5.1}$$

where NO_{2_i} is the averaged annual concentration for data zone i. Here $\tilde{\mathrm{NO}}_{21}, \ldots, \tilde{\mathrm{NO}}_{2n_i}$ are the modelled NO_2 concentrations at the n_i grid squares within data zone i, and d_{ij} is the Euclidean distance between the population-weighted centroid of data zone i and the centroid of grid square j. This is a distance weighting formula that gives more weight to grid squares closer to the population weighted centroid compared to further away grid squares. This ensures that the data zone takes a representative value according to the location at which the population density is greatest, as dictated by the population-weighted centroid. The aggregation approach is based on three typical scenarios (examples of which are displayed in Figure 5.2): data zones containing no grid square centroids, data zones containing only one grid square centroid, and data zones containing more than one grid square centroid. Data zones containing no grid square centroids (see Figure 5.2a) were assigned the NO_2 concentration nearest the population-weighted centroid of the data zone. These data zones typically occur in the most populated urban environments in Glasgow, where the data zones are extremely small due to the high population density. Data zones containing only one grid square centroid are assigned the NO_2 concentration of that grid square (see Figure 5.2b), and data zones containing more than one grid square centroid (see Figure 5.2c) are aggregated based on the above formula.

The data zone averaged modelled concentrations are displayed in the top left panel

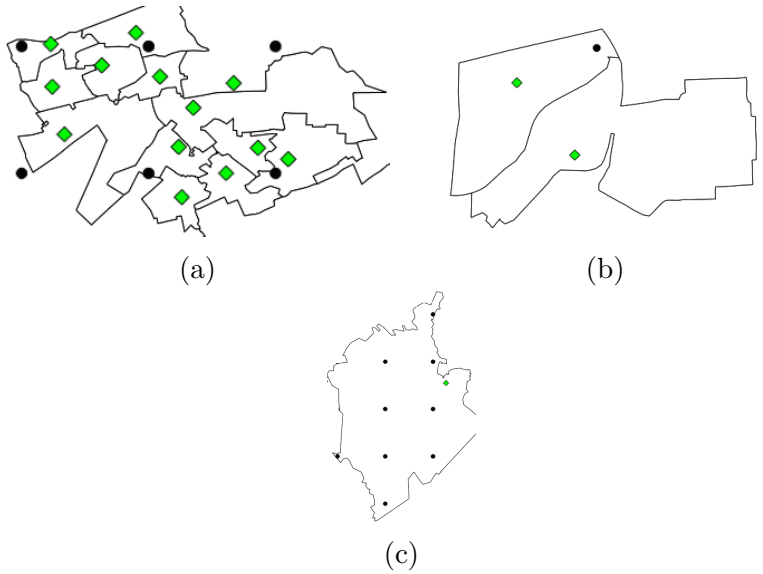

Figure 5.2: *Three data zone scenarios for aggregating NO_2 pollutant concentrations. Dots represent the grid box centroids and diamonds represent the population-weighted centroids. (a) depicts scenario one, where there are data zones that contain no grid box centroid. (b) depicts scenario two, where there are data zones that contain only one grid box centroid. (c) depicts scenario three, where a data zone contains more than one grid box centroid.*

of Figure 5.1. As expected, background concentrations are highest for the city of Glasgow. However, these modelled concentrations are known to contain biases, and so this chapter compares the health effects estimated from using them to those obtained from predicting NO_2 with a statistical fusion model.

The second set of NO_2 concentrations were developed in Chapter 4, specifically for the West Central Scotland study region. The general form of the model is given by:

$$\begin{aligned} Z(\mathbf{s}_i) &\sim \mathrm{N}\Big(\mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \phi(\mathbf{s}_i), \nu^2\sigma^2\Big), \quad i = 1, \ldots, m, \\ \boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_m)) &\sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}(\rho)), \end{aligned} \tag{5.2}$$

where $Z(\mathbf{s}_i)$ is the measured NO_2 concentration at spatial location $\mathbf{s}_i$ for $i = 1, \ldots, m$ spatial locations. The n measurements are modelled by a set of covariates $\mathbf{x}(\mathbf{s}_i)$ with regression parameters, $\boldsymbol{\beta}$, and the former include the modelled concentration in the nearest grid square, the local environment in which the site is located (e.g. roadside, urban background, rural), and an interaction term between the modelled concentration and the spatial location of the measured data. The second term in the mean function is a vector of spatial random effects, $\boldsymbol{\phi} = (\phi(\mathbf{s}_1), \ldots, \phi(\mathbf{s}_n))$, which accounts for residual spatial autocorrelation in the measured data, and is modelled by a Gaussian process with a spatial exponential correlation matrix, $\boldsymbol{\Sigma}(\rho)$, and range parameter ρ. Full details of this model are discussed in Chapter 4.

As West Central Scotland has a large proportion of rural areas, this model calibrates the modelled concentrations according to urban background and rural environments, while allowing the effect of the modelled concentrations to vary linearly across space. The predictions were made from the model on a 1km $\times$ 1km resolution for each year separately. These predictions were temporally aggregated by averaging and then spatially aggregated to data zone level using the same form as Equation (5.1). The resulting concentrations are displayed in the top right panel of Figure 5.1, which is structurally similar to the modelled concentrations as expected.

5.2.3 Deprivation data

Many studies have shown that populations living in more deprived areas exhibit greater levels of morbidity and mortality compared to populations living in more affluent areas (Mackenbach et al., 2008; Smith et al., 1990). Therefore, the main confounding factor in ecological health studies is socio-economic deprivation (Mackenbach et al., 1997), in which populations with higher levels of deprivation may be more susceptible to the effects of air pollution (Laurent et al., 2007; O'Neill et al., 2003). This may be due to individuals living in more deprived communities having worse underlying health, on average, than those living in more affluent communities. The majority of the most deprived areas in Scotland occur within Glasgow, and Scotland is infamous for its low life expectancy compared to other Western European countries (McCartney et al., 2012; Schofield et al., 2016). Deprivation is a known determinant of health and is paramount when assessing the relationship between air pollution and ill health, not just in West Central Scotland, but in any study region. However, deprivation is multi-factorial and difficult to measure, but is commonly represented by a composite index. This chapter makes use of the Scottish Index of Multiple Deprivation (SIMD, `http://www.gov.scot/Topics/Statistics/SIMD`), which is a composite index containing seven domains, namely: access to services; crime; education, skills and training; employment; income; health; and housing. However, as the health domain contains an indicator of the Comparative Mortality Factor, and as such, includes deaths which are part of the outcome in this chapter, it has not been included here. The 2009 version of the index was selected because it forms the mid-point of the study period.

The domains are available as continuous measures comprising scores, rates and counts. Continuous measures for the education, skills and training domain; housing domain; and geographic access to services domain are based on scores, whereby the individual indicators are ranked, transformed to a standard normal, and then combined using weights generated by a factor analysis. The income and employment domains have continuous measures based on rates. For the income domain, it represents the percentage in each data zone who are in receipt of benefits, such as income support. For the employment domain, it represents the percentage of each data zone's working

age population who are in receipt of benefits, such as unemployment claimant count, working age incapacity benefit, and employment support allowance. The crime domain comprises a count relating to selected recorded offences in the data zone, rather than all crimes committed in the area. The selected crimes relate to violence, domestic housebreaking, vandalism, drugs offences and minor assaults. The mean number of crimes committed is 45.266, with a standard deviation of 67.044, suggesting a wide range of values for this variable. Correlations between the remaining six domains are displayed in Table 5.2, where there are high correlations between income, employment, and education; and weak to moderate correlations with the access, housing and crime domains. Access has considerably lower correlations with all other domains suggesting it is exhibiting an independent spatial pattern.

Table 5.2: *Correlations between the six deprivation measures, where EST denotes the education, skills and training domain.*

Variable	Access	Crime	EST	Employment	Income	Housing
Access	1	-0.252	-0.250	-0.287	-0.321	-0.411
Crime	-0.252	1	0.411	0.436	0.430	0.351
EST	-0.250	0.411	1	0.833	0.860	0.680
Employment	-0.287	0.436	0.833	1	0.946	0.436
Income	-0.321	0.430	0.860	0.946	1	0.658
Housing	-0.411	0.351	0.680	0.436	0.658	1

The SIMD also comprises an overall score, which is a weighted sum of the seven domains. As this also includes the health domain, it is not appropriate to be used as a covariate. Therefore, the overall score was re-weighted to remove the health domain and was constructed based on the original index methodology (http://www.gov.scot/Publications/2004/10/20089/45173). Briefly, the new overall score was constructed by transforming the ranks of the individual domains to an exponential distribution using the formula

$$T = -23 \times \log[1 - R(1 - \exp(-100/23))], \tag{5.3}$$

where R denotes the rank of the domain transformed to the range $[0, 1]$. The domains are ranked to standardise them since they are on different scales. This ensures they have identical distributions with the same range. However, the ranks result in distributions that are symmetrical, where it is possible that high levels of deprivation in one domain cancel out low deprivation levels in another domain. Using the exponential transformation mitigates against this. The constant in equation (5.3) -23 gives a 10% cancellation property so that data zones are ranked within the 10% most deprived data zones. The domains then have scores that range from 0 (least deprived) to 100 (most deprived), and are combined into an overall score by summing the scores to the following weights: 12 for income, 12 for employment, 6 for education, 4 for access, 2 for

crime and 1 for housing. The bottom right panel of Figure 5.1 displays the re-weighted overall score, where it is clear that the City of Glasgow contains the majority of the most deprived areas, as expected.

Initially, a simple Quasi-Poisson generalised linear model (without any spatial random effects) was applied to the disease data, with NO_2 (DEFRA modelled concentrations) and income deprivation as covariates. The overdispersion parameter was estimated as 4.35, suggesting substantial overdispersion with respect to the Poisson assumption of equal mean and variance. The residuals from this model were tested for any residual spatial autocorrelation, using a permutation test based on Moran's I statistic (Moran, 1950) given by equation (2.76) in Chapter 2 Section 2.4.2. The null hypothesis of this test is no spatial autocorrelation, and Moran's I statistic was 0.036 with a p-value of 0.003, suggesting that spatial autocorrelation is present in the residuals. Since spatial autocorrelation is present in the data, spatial models are therefore required to take this residual spatial autocorrelation into account. There are numerous ways of modelling this residual spatial autocorrelation, of which three approaches are described below.

5.3 Statistical models for estimating air pollution and health effects

The aim of this chapter is to estimate the sensitivity of the estimated relationship between NO_2 concentrations and cardio-respiratory mortality in the West Central Scotland region between 2006 and 2012. This included estimating the sensitivity of the pollutant-health effect to changing the estimation of NO_2 concentrations, controlling for socio-economic deprivation, and allowance for residual spatial autocorrelation in the mortality data after adjusting for the covariate effects. Three specific Poisson log-linear models are compared in this sensitivity analysis, which differ in their control for residual spatial autocorrelation. These models are then combined to estimate an overall pollution-health effect using Bayesian model averaging, with inference based on Markov chain Monte Carlo (McMC) simulation. These models are implemented in the `R` software environment (R Core Team, 2015), using self-written code, the `CARBayes` (Lee, 2013) and `ngspatial` (Hughes & Cui, 1-16-2015) packages. Sensitivity to the estimation of NO_2 and control for socio-economic deprivation is assessed by fitting different covariate combinations in all of the three models described below.

5.3.1 Data and Likelihood model

The vector of the observed numbers of cardio-respiratory deaths is denoted by $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$, while the expected numbers of deaths are computed using indirect

standardisation (discussed in Chapter 3, Section 2.5) based on age- and sex-specific cardio-respiratory mortality rates in West Central Scotland. These expected counts are denoted by $\mathbf{E} = (E_1, \ldots, E_m)^\top$, where for data zone i, $E_i = \sum_r N_{ir}\gamma_r$, where N_{ir} is the number of people in age-sex group r in data zone i, and γ_r denotes the region-wide age-sex mortality rate. The vector of NO_2 concentrations is denoted by $\mathbf{x} = (x_1, \ldots, x_m)$ for all m data zones, while each measure of socio-economic deprivation is denoted by $\mathbf{u} = (u_1, \ldots, u_m)$. Thus, for the ith data zone, the vector of covariates is given by $\mathbf{z}_i^\top = (1, x_i, u_i)$, while the corresponding regression parameters are given by $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$, so that β_1 is the intercept term, and β_2 is the key parameter in this model, namely the effect of NO_2 on cardio-respiratory mortality risk. A general Bayesian Poisson log-linear model for these data is given by:

$$\begin{aligned} Y_i \mid E_i, R_i &\sim \text{Poisson}(E_i R_i) \quad \text{for } i = 1, \ldots, m, \\ \ln(R_i) &= \mathbf{z}_i^\top \boldsymbol{\beta} + \phi_i, \\ \boldsymbol{\beta} &\sim \text{N}(\mathbf{m}, \mathbf{V}), \end{aligned} \tag{5.4}$$

where R_i is the risk of disease in data zone i. The regression parameters, $\boldsymbol{\beta}$, are assigned a weakly informative multivariate Gaussian prior, with hyperparameters $(\mathbf{m}, \mathbf{V})$, typically with mean zero and a large diagonal variance matrix. Population demography, including the age-sex structure of the population and the overall size, can pose some confounding effects and are partially accounted for within the expected counts of mortality (E_i), by including it as an offset in the regression models. The final term in the linear predictor is the vector of random effects, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$, which controls the residual spatial autocorrelation in the data after accounting for the covariate effects. Three modelling specifications are now considered here, which include ignoring the presence of any residual spatial autocorrelation, utilising a commonly-used approach to modelling residual spatial autocorrelation, and utilising an approach that tries to alleviate the issues surrounding the commonly-used approaches.

5.3.2 Model 1 - no spatial autocorrelation

The simplest approach is to ignore the presence of any residual spatial autocorrelation and assume $\phi_i = 0$ for all data zones, i, which is equivalent to fitting a Poisson generalised linear model to the data. This model naively assumes the cardio-respiratory counts are independent conditional on the covariates, which, as illustrated in Section 5.2.3, is not true for this case study. Additionally, the model does not allow for overdispersion relative to the Poisson likelihood, and thus makes the restrictive assumption that $\mathbb{E}[Y_i] = \text{Var}[Y_i]$ (see Chapter 2 Section 2.2.1), which is unrealistic. This model is thus only included here for comparison purposes, with the remaining two models being described below.

5.3.3 Model 2 - globally smooth spatial autocorrelation

The standard approach to accounting for residual spatial autocorrelation and overdispersion in this context is to model $\boldsymbol{\phi}$ by a set of globally spatially smooth (autocorrelated) random effects, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$. A number of models can be specified for these random effects, including conditional autoregressive (CAR), simultaneous autoregressive (SAR) or geostatistical models. However, CAR priors are the most common in this field, and examples of their use include Maheswaran et al. (2005a) and Lee et al. (2009). A number of globally smooth CAR priors have been proposed, and a review by Lee (2011) concluded that the model proposed by Leroux et al. (1999) was the most appealing because its results were consistent across a range of spatial autocorrelation scenarios, and it is flexible in its ability to account for both strong and weak spatial autocorrelation structures. Furthermore, the strength of the spatial autocorrelation is captured within one set of random effects, which makes it superior compared to other CAR specifications (see Chapter 2 Section 2.4.2.3) that do not encompass this component. In addition, its specification corresponds to a proper joint distribution for the random effects, and it does not assume the conditional variance is inversely proportional to the total number of neighbouring areas, even when no spatial autocorrelation is present, therefore adding to its flexibility. This model can be specified by a set of m univariate full conditional distributions, $p(\phi_i \mid \boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_m)$. Spatial autocorrelation is imposed using a binary $m \times m$ neighbourhood matrix, $\mathbf{W}$, whose ijth element $w_{ij} = 1$ if areas (i, j) share a common border, and $w_{ij} = 0$ otherwise. This specification asserts that neighbouring areas have random effects that are partially autocorrelated, otherwise the random effects are conditionally independent. The model has the form

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\rho \sum_{j=1}^{m} w_{ij}\phi_j}{\rho \sum_{j=1}^{m} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{m} w_{ij} + 1 - \rho}\right), \tag{5.5}$$

where ρ controls the level of spatial autocorrelation. Spatial independence occurs when $\rho = 0$, with mean zero and constant variance. Strong spatial autocorrelation is defined when $\rho = 1$, which simplifies to the intrinsic CAR model given by equation (2.84) in Chapter 2 Section 2.4.2.1. Weakly informative hyperpriors are assigned for τ^2 and ρ; typically an inverse-gamma(a, b) distribution for τ^2, and a uniform distribution on the unit interval for ρ.

5.3.4 Model 3 - orthogonal smoothing

One problem with traditional CAR models, such as (5.5), is that the spatially smooth random effects have been shown to be correlated with the spatially smooth covariates, such as air pollution (Clayton et al., 1993; Paciorek, 2010). This spatial confounding between the fixed and random effects leads to variance inflation and the model param-

eters becoming uninterpretable. Much research has been conducted on controlling for this spatial confounding, in which the random effects are instead modelled with a series of basis functions that are orthogonal to the covariates, thus mitigating this confounding (Hughes & Haran, 2013; Reich et al., 2006). This chapter utilises the orthogonal smoothing model proposed by Hughes & Haran (2013) due to the low dimensionality of the random effects, which leads to fast computation. This model replaces the vector of random effects ϕ_i in equation (5.4) with a linear combination of basis functions that are orthogonal to the fixed effects.

Let the matrix of p covariates be denoted by $\mathbf{Z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_m^\top)^\top$. Then the orthogonal projection matrix (hat matrix) onto the column space of the design matrix $\mathbf{Z}$ is defined by

$$\mathbf{P} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top. \tag{5.6}$$

Further, let the residual projection matrix onto the space orthogonal to $\mathbf{Z}$ be defined by

$$\mathbf{P}' = \mathbf{I}_m - \mathbf{P}. \tag{5.7}$$

The residual projection matrix is then used to create a set of eigenvectors, given by the matrix product, $\mathbf{P}'\mathbf{W}\mathbf{P}'$, which combines covariate orthogonality given by $\mathbf{P}'$ with spatial adjacency given by $\mathbf{W}$. The eigenvectors of $\mathbf{P}'\mathbf{W}\mathbf{P}'$ contain all possible mutually distinct spatial patterns of clustering orthogonal to $\mathbf{Z}$. Furthermore, spatial dependence is related to both positive and negative eigenvalues, where positive eigenvalues correspond to positive spatial autocorrelation. The size of the eigenvalue associated with a given eigenvector determines the relative importance of its spatial pattern, so Hughes & Haran (2013) suggest only selecting the first $q << m$ eigenvectors corresponding to the largest positive eigenvectors. This matrix is denoted by $\mathbf{B}$, where $\mathbf{b}_i^\top = (b_{i1}, \ldots, b_{iq})$. The chosen number of eigenvectors, q, acts as a tuning parameter, which determines the extent of dimensionality reduction in the model. The authors suggest using $q = 50$ as a default choice. The orthogonal smoothing model replaces the random effects in the linear predictor in equation (5.4) by

$$\begin{aligned} \ln(R_i) &= \mathbf{z}_i^\top\boldsymbol{\beta} + \mathbf{b}_i^\top\boldsymbol{\delta}, \\ \boldsymbol{\delta} &\sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W})_s^{-1}), \end{aligned} \tag{5.8}$$

where the random effects, $\boldsymbol{\delta}$, are assigned a Gaussian prior with mean $\mathbf{0}$, and precision matrix given by $\mathbf{Q}(\mathbf{W})_s = \mathbf{B}^\top\mathbf{Q}(\mathbf{W})\mathbf{B}$, where $\mathbf{Q}(\mathbf{W}) = \mathrm{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$ corresponds to the precision matrix for the intrinsic CAR prior (Besag et al., 1991) given by equation (2.84).

5.3.5 Bayesian model averaging

Bayesian model averaging provides a coherent framework for combining estimates of the same quantity of interest from a number of different Bayesian models into a single overall estimate to account for model uncertainty. Such model uncertainty is often ignored in existing studies, and as shown in the next section, can have a large impact on the results. Recall that β_2 is the key parameter of interest in this model, namely the effect of NO_2 concentrations on cardio-respiratory mortality risk. Consider the case of having K candidate models, where in this chapter there are $K = 42$ models (see Section 5.4 for details). Denote these models by $(M_1, \ldots, M_K)$ and their respective sets of model parameters by $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Let β_2 denote the true unknown parameter of interest and $\hat{\beta}_{2_k}$ denote the estimate (posterior median) from the kth model. Then the posterior distribution of interest is

$$p(\beta_2 \,|\, \mathbf{Y}) = \sum_{k=1}^{K} p(\beta_2 \,|\, M_k, \mathbf{Y}) p(M_k \,|\, \mathbf{Y}). \tag{5.9}$$

Here, $p(\beta_2 \,|\, M_k, \mathbf{Y})$ is the posterior distribution of β_2 from model K, and $p(M_k \,|\, \mathbf{Y})$ is the posterior probability of model M_k. This equation essentially averages the posterior distributions for NO_2 under each model weighted by their posterior model probabilities.

The posterior probability for model M_k is given by

$$p(M_k \,|\, \mathbf{Y}) = \frac{p(\mathbf{Y} \,|\, M_k) p(M_k)}{\sum_{l=1}^{K} p(\mathbf{Y} \,|\, M_l) p(M_l)}, \tag{5.10}$$

where $p(M_k)$ is the prior probability for model M_k. Prior ignorance is specified via a discrete uniform prior for $p(M_k)$, that is $p(M_k) = 1/K$. This specification simplifies the posterior probability for model M_k in equation (5.10) to

$$p(M_k \,|\, \mathbf{Y}) = \frac{p(\mathbf{Y} \,|\, M_k)}{\sum_{l=1}^{K} p(\mathbf{Y} \,|\, M_l)}. \tag{5.11}$$

The marginal (averaged over the parameters) probability of the data given model M_k is computed by

$$p(\mathbf{Y} \,|\, M_k) = \int_{\boldsymbol{\theta}_k} p(\mathbf{Y} \,|\, \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k \,|\, M_k) \mathbf{d}\boldsymbol{\theta}_k, \tag{5.12}$$

which can be approximated by J McMC samples as

$$p(\mathbf{Y} \,|\, M_k) \approx \frac{1}{J} \sum_{j=1}^{J} p(\mathbf{Y} \,|\, \boldsymbol{\theta}_k^{(j)}, M_k) p(\boldsymbol{\theta}_k^{(j)} \,|\, M_k), \tag{5.13}$$

where $\boldsymbol{\theta}_k^{(j)}$ is the jth McMC sample for model M_k. Once these quantities have been computed, the posterior mean and variance of β_2 are given by:

$$\mathbb{E}[\beta_2 \,|\, \mathbf{Y}] = \sum_{k=1}^{K} \hat{\beta}_{2_k} p(M_k \,|\, \mathbf{Y}), \tag{5.14}$$

$$\text{Var}[\beta_2 \,|\, \mathbf{Y}] = \sum_{k=1}^{K} (\text{Var}[\beta_2 \,|\, M_k, \mathbf{Y}] + \hat{\beta}_{2_k}^2) p(M_k \,|\, \mathbf{Y}) - \mathbb{E}[\beta_2 \,|\, \mathbf{Y}]^2, \tag{5.15}$$

where $\text{Var}[\beta_2 \,|\, M_k, \mathbf{Y}]$ is the posterior variance of β_2 from model M_k. Based on a normal approximation to the posterior, an approximate 95% credible interval (CI) can be obtained for β_2 that accounts for model uncertainty as follows:

$$95\% \text{ CI} = \mathbb{E}[\beta_2 \,|\, \mathbf{Y}] \pm 1.96\sqrt{\text{Var}[\beta_2 \,|\, \mathbf{Y}]}. \tag{5.16}$$

5.4 Results from the West Central Scotland study

This section presents results from investigating the long-term effects of NO_2 concentrations on cardio-respiratory mortality in West Central Scotland between 2006 and 2012 inclusive. Section 5.4.1 describes the set of results obtained from fitting the statistical models described in Section 5.3, which illustrates the sensitivity of the results to model choice. Section 5.4.2 presents the overall estimate of the effect of NO_2 on cardio-respiratory mortality using the Bayesian model averaging approach, as outlined in the previous section. Inference for all models described in this section are based on running 5 parallel Markov chains for 120,000 iterations, which included a burn-in period of 20,000 iterations. The remaining samples were thinned by 10 to reduce their autocorrelation, thus producing a final set of 50,000 posterior samples across the five chains.

5.4.1 Results - sensitivity to model choice

This section empirically investigates the sensitivity of the estimated pollution-health effect to three modelling choices. The first is the estimation of spatially averaged NO_2 concentrations for each data zone, and compares averaging the raw output from the atmospheric dispersion model used by DEFRA (`http://uk-air.defra.gov.uk/`, denoted *DEFRA*) to averaging predictions from the fusion model proposed in Chapter 4 (denoted *Fusion*). The second modelling choice concerns control for the confounding effects of socio-economic deprivation, and compares using the composite Scottish Index of Multiple Deprivation (SIMD, minus the health domain), with individual indicators from its sub-domains, namely access to services, crime, education, employment, housing and income. Finally, three approaches to controlling for residual spatial autocorrelation are compared: ignoring it (denoted *GLM*), modelling it using random effects represented by the globally smooth model proposed by Leroux et al. (1999) (denoted *Leroux*), and modelling it using a set of orthogonal random effects proposed

by Hughes & Haran (2013) (denoted *OS*).

All combinations of these factors give a set of 42 possible models, and the results are summarised in Tables 5.3 and 5.4, which respectively display the posterior median relative risks and 95% credible intervals, and the Deviance Information Criterion (DIC, given by equation (2.37)) together with the effective number of parameters (p_D) for each model, which were described in Chapter 2 Section 2.3.4. All pollution-health effects are presented on the relative risk scale for a $5\mu gm^{-3}$ increase in NO_2 concentrations (for both *DEFRA* and *Fusion*), as this is a realistic change in long-term exposure.

Overall, there is evidence that increasing NO_2 concentrations are associated with small but positive increases in the risk of cardio-respiratory mortality, as 36 out of the 42 models estimate the relative risk to be greater than 1. However, the range of the effects estimated across the 42 models is large, being between a 2% decreased risk (0.980) to a 5.3% increased risk (1.053) associated with a $5\mu gm^{-3}$ increase in NO_2. This suggests that the results are highly sensitive to model choice, and that if results were presented from a single model then either a positive or a negative effect of NO_2 on mortality risk could have been observed. Focusing on the 95% credible intervals shows that 23 of the 42 intervals are wholly above the null risk of 1, which is just over 57% of the models considered.

The three modelling choices considered here all appear to have the potential to substantially affect the estimated relative risks, as the estimates from varying one factor at a time can lead to large changes in risk. For example, changing the NO_2 metric from that produced by the fusion model to that produced by DEFRA resulted in the risks changing by between -1% and 5.3%, and in all but 3 cases, these changes were positive. This indicates that, overall, using the DEFRA concentrations resulted in increased risks compared with using the predictions from the fusion model.

Changing how socio-economic deprivation was controlled for also had a large impact on the results, with changes in relative risk of between 3.8% and 7.3% across the 7 measures considered depending on the combination of *DEFRA/Fusion* and *GLM/Leroux/OS*. In general, using the housing indicator resulted in the lowest effect sizes, while using crime resulted in the highest estimates. Finally, varying the control for spatial autocorrelation had a slight effect on the results, with differences in relative risk between the three models considered ranging between 0.1% and 0.9%. The only pattern of note is that the effect sizes are attenuated for the *OS* models compared to the *Leroux* models in 10 out of the 14 models, with 2 models comprising the same effect size.

Finally, Table 5.4 summarises the fit of each model via the DIC, which shows that,

Table 5.3: *Posterior median relative risks (RR) and 95% credible intervals for a 5μgm^{-3} increase in NO_2 concentrations on cardio-respiratory mortality. The results displayed relate to models varying in their estimation of NO_2, control for deprivation and allowance for residual spatial autocorrelation. The results in bold are substantial effects at the 5% level.*

Deprivation	Model	RR (95% CI)	
		Fusion	DEFRA
Access	GLM	**1.036 (1.016, 1.056)**	**1.050 (1.026, 1.075)**
	Leroux	**1.033 (1.006, 1.059)**	**1.045 (1.015, 1.075)**
	OS	**1.029 (1.020, 1.039)**	**1.041 (1.030, 1.051)**
Crime	GLM	**1.038 (1.018, 1.058)**	**1.053 (1.033, 1.074)**
	Leroux	**1.039 (1.015, 1.063)**	**1.053 (1.027, 1.079)**
	OS	**1.034 (1.025, 1.043)**	**1.046 (1.037, 1.057)**
Education	GLM	1.006 (0.988, 1.024)	1.019 (0.999, 1.039)
	Leroux	1.007 (0.991, 1.024)	**1.019 (1.002, 1.041)**
	OS	1.006 (0.998, 1.015)	**1.021 (1.011, 1.030)**
Employment	GLM	1.010 (0.990, 1.030)	**1.020 (1.000, 1.040)**
	Leroux	1.015 (0.998, 1.033)	**1.025 (1.007, 1.044)**
	OS	**1.014 (1.006, 1.023)**	**1.025 (1.016, 1.036)**
Housing	GLM	0.992 (0.973, 1.012)	0.989 (0.968, 1.011)
	Leroux	0.990 (0.971, 1.009)	0.980 (0.959, 1.002)
	OS	0.992 (0.983, 1.002)	**0.987 (0.977, 0.997)**
Income	GLM	1.003 (0.985, 1.021)	1.010 (0.990, 1.030)
	Leroux	1.008 (0.992, 1.018)	1.012 (0.995, 1.030)
	OS	1.007 (0.998, 1.015)	**1.013 (1.004, 1.023)**
SIMD	GLM	1.007 (0.989, 1.025)	1.017 (0.997, 1.037)
	Leroux	1.013 (0.997, 1.030)	**1.021 (1.003, 1.040)**
	OS	**1.011 (1.003, 1.020)**	**1.021 (1.011, 1.030)**

in all cases, the *Leroux* model fits the data best compared with the other alternatives. This is surprising considering the globally smooth model has the potential for correlation between the fixed and random effects and thus one would expect the *OS* model to outperform the *Leroux* model. The *OS* model has many fewer effective numbers of parameters compared to the *Leroux* model, which makes it more parsimonious. However, it is this reduction in dimensionality that has resulted in a poorer fit to the data (in terms of DIC). In most cases, the DIC is lower for the *DEFRA* concentrations compared to the *Fusion* concentrations, while the income domain provides the best fit to the data of all the socio-economic indicators considered here. Furthermore, the RMSE allows the closeness of the models fitted values to the observed health outcomes (with the lowest values indicating better performance) to be assessed, and to give a sense of scale, the 25th and 75th percentiles of the observed cardio-respiratory deaths were 14 and 33 respectively. The *Leroux* models have the lowest RMSE values compared to the *GLM*s and *OS* models, which is due to their increased number of effective parameters. For the *Leroux* models the relationship between RMSE and deprivation is opposite to that observed for the *GLM*s and *OS* models, with income, the best deprivation co-

variate in terms of DIC, having the highest RMSE compared to the other deprivation measures (2.693 compared to 2.518 for access). This can be explained by the lower residual spatial variation in the model adjusting for income deprivation compared to using the other deprivation covariates. The random effects have less spatial variation, and thus, less impact on the fitted values. For example, the variance, τ^2, is 0.230 for the access covariate compared to 0.094 for income. This is also observed in the effective number of parameters p_D, which is smallest for the model with income. In contrast, the DIC is an overall measure of model quality that penalises complex models that contain more parameters, as is observed in higher DIC values for access (and others) compared to income.

Table 5.4: *Model fit for each of the 42 models, measured by the Deviance Information Criterion (DIC), the effective number of parameters (p_D), and the root mean square error (RMSE)*

Deprivation	Model	DIC (p_D)		RMSE	
		Fusion	DEFRA	Fusion	DEFRA
Access	GLM	20219 (2)	20182 (2)	13.560	13.519
	Leroux	13797 (1508)	13799 (1507)	2.518	2.525
	OS	19130 (76)	19115 (74)	12.614	12.604
Crime	GLM	20017 (2)	19967 (2)	13.471	13.429
	Leroux	13793 (1498)	13791 (1497)	2.511	2.510
	OS	19222 (67)	19201 (66)	12.707	12.697
Education	GLM	18240 (2)	18224 (2)	12.742	12.724
	Leroux	13601 (1369)	13600 (1367)	2.687	2.688
	OS	17964 (62)	17942 (62)	12.336	12.319
Employment	GLM	18373 (2)	18359 (2)	12.811	12.812
	Leroux	13600 (1378)	13597 (1377)	2.655	2.658
	OS	18010 (66)	17996 (65)	12.323	12.318
Housing	GLM	19107 (2)	19106 (2)	12.989	12.993
	Leroux	13737 (1451)	13736 (1450)	2.522	2.522
	OS	18336 (69)	18352 (69)	12.302	12.323
Income	GLM	18139 (2)	18135 (2)	12.638	12.623
	Leroux	13589 (1362)	13589 (1362)	2.693	2.692
	OS	17743 (66)	17729 (60)	12.128	12.129
SIMD	GLM	18277 (2)	18267 (2)	12.701	12.694
	Leroux	13609 (1374)	13606 (1373)	2.672	2.670
	OS	17900 (62)	17898 (64)	12.242	12.240

5.4.2 Results - BMA

The previous section shows clear sensitivity of the results to the model fitted, and one solution would be to choose a single 'best' model, for example, by minimising the DIC. However, this clearly ignores model uncertainty, which can be accounted for using BMA as described in Section 5.3.5. This method combines the estimated effect sizes from the 42 models considered here. When this was conducted, the overall estimated relative risk was 1.011 together with an associated 95% uncertainty interval of (0.993,

1.029). This small, but positive effect indicates that for a $5\mu gm^{-3}$ increase in NO_2 concentrations, cardio-respiratory deaths increase by an estimated 1.1%, although it should be noted that the lower end of the 95% credible interval is below the null risk of 1. In fact, the posterior probability that the relative risk is greater than 1 is 0.884. This result is essentially a mixture of the effect estimates from the *Leroux* model including income and *DEFRA* NO_2 concentrations, and the effect estimate from the *Leroux* model including income and *Fusion* model NO_2 concentrations. The former had the most influence on the overall effect size, since its posterior model probability, $p(M_k|\,\mathbf{Y})$, was 67.82%, whilst it was 32.17% for the latter. So, in this example, the large differences in fit across the 42 models has resulted in only two models contributing to the overall effect estimate.

5.5 Discussion

In this chapter, sensitivity of the pollution-health relationship in West Central Scotland to the impact of three modelling choices was investigated: the estimation of NO_2 concentrations, control for socio-economic deprivation, and control for residual spatial autocorrelation after accounting for covariate effects. The main finding is that the choice of these three factors can have a major impact on the resulting pollution-health effects, meaning that presenting results from a single model could result in a wide range of effect sizes depending on the model selected. The estimated pollution-health effect in this study varies considerably across the 42 models (effect sizes range from 0.980-1.053), highlighting the estimated pollution-health effect sizes are not robust to the three aforementioned factors.

BMA was utilised to combine results from all 42 models into an overall pollution-health effect size, whilst taking model uncertainty into account. The final estimated effect size shows that a $5\mu gm^{-3}$ increase in NO_2 concentrations is associated with 1.1% higher cardio-respiratory deaths in West Central Scotland between 2006 and 2012. However, this effect is (borderline) not substantial at the 5% level, as the resulting 95% credible interval contains the null risk of 1. This could be due to the fact that the majority of NO_2 concentrations are relatively low, and thus greater variation in the exposure would be needed to observe substantial health impacts.

A second finding is the attenuation of the pollution-health effects when the NO_2 concentrations were estimated using the geostatistical fusion model proposed in Chapter 4, compared to when the NO_2 concentrations were estimated by the DEFRA diffusion model. The estimated health effects changed between -1% and 5.3%, indicating that increased risks are observed when the DEFRA concentrations are utilised. This is an interesting result considering that the majority of spatial ecological studies in Scotland, and indeed in the UK, make use of modelled concentrations due its wide availability and

fine scale spatial coverage. Furthermore, the correlation between residual disease (after adjustment from income deprivation) and pollution from both the fusion and DEFRA models is 0.041 and 0.029 respectively. This highlights that the DEFRA pollution concentrations are more correlated with residual disease, thus explaining its stronger effect size (see Table 5.3) compared to the pollution concentrations from the fusion model. However, in terms of pollution predictive performance, Section 4.4.1, shows that the DEFRA data are not as good at predicting measured pollution concentrations at the point level, since the root mean square prediction error (RMSPE) is 0.337 compared to 0.255 for the fusion model. A recent study conducted in mainland Scotland by Huang et al. (2015), concluded that the estimated health effects of NO_2 were largely consistent when estimated from a fusion model compared to modelled concentrations from the DEFRA. However, the authors utilised a coarser spatial resolution compared to this chapter, suggesting further research is needed to understand why changing spatial resolution changes the results of the estimated NO_2-health relationship when different types of NO_2 concentrations are used.

A third finding is that the global spatial autocorrelation model comprising the DEFRA concentrations and income deprivation dominated the overall pollution-health effect size when combining models using BMA. The posterior model probability was 67.82%, while for the model with the fusion model concentrations it was 32.17%. It is interesting to note that only 2 out of the 42 models had a considerable influence, suggesting that the global spatial autocorrelation model, income deprivation and DEFRA concentrations are the most important factors when investigating the impact of air pollution on health in West Central Scotland. In addition, the global spatial autocorrelation model, which has been under much scrutiny by Reich et al. (2006), outperformed the orthogonal smoothing model proposed by Hughes & Haran (2013) in terms of model fit via the DIC.

However, here are a few limitations to these analyses. Firstly, cardio-respiratory deaths were aggregated over a 7 year period to ensure sufficient variation in the disease data. This meant it was not possible to investigate how pollution-health risks had changed over time, since the analyses were purely spatial. In addition, small numbers of events from cardio-respiratory deaths at the data zone level may necessitate the need to upgrade to a larger spatial resolution, such as intermediate geographies, comprising 4300 inhabitants on average, in order to improve the power to detect an association. Alternatively, hospital admission data could instead be used to create the health outcome of interest, as it is expected that there will be more events and thus, no need to aggregate over multiple years.

Secondly, the DEFRA modelled concentrations come with no measure of uncertainty, which could impact the analysis. The predicted concentrations from the geosta-

tistical fusion model do have measures of prediction uncertainty; however, in this study the predicted NO_2 concentrations were treated as the known and true values, which again could impact results. Therefore, an avenue for future work is to incorporate the uncertainty surrounding predicted NO_2 concentrations in an combined Bayesian framework, which estimates the exposures and health risks simultaneously.

Chapter 6

Investigating the long-term effect of outdoor air pollution on cardio-respiratory incidence in West Central Scotland

6.1 Introduction

The previous chapter explored the relationship between air pollution and cardio-respiratory deaths by applying predicted NO_2 concentrations, obtained from a novel statistical fusion model, to mortality data. This chapter will instead focus on the incidence of cardio-respiratory disease as the health outcome. The outcome was created by combining cardio-respiratory deaths with first-ever admission to hospital for causes related to cardio-respiratory disease as the primary diagnosis. These data are the first of their kind to be used in an air pollution and health study in Scotland. Furthermore, no previous research in Scotland has focussed on incidence.

Age is an important consideration when it comes to making policy decisions regarding air quality, as more vulnerable groups, such as children and the elderly, may be more susceptible to the adverse effects of air pollution. Young children can be more prone to chronic respiratory conditions, like asthma, and older people are more likely to have numerous co-morbidities, and are also just generally more susceptible because of old age (Beatty & Shimshack, 2014). Thus, the effect of NO_2 on the risk of cardio-respiratory ill health will be studied at different age groups. A more detailed discussion of age with respect to air pollution and health studies can be found in Chapter 3 Section 3.3.2. This is the first epidemiological study in Scotland to investigate the pollution-health relationship across the age spectrum as part of a spatial ecological study.

In addition to age being an important variable to consider, it is well acknowledged in the air pollution and ill health literature that deprivation plays a key role in the estimation of the relationship between air pollution and ill health. Populations residing in more deprived areas may be at an increased risk of pollutant-related morbidity or mortality (Carder et al., 2010), or exacerbating certain medical conditions, such as asthma. It is thought that more deprived people may reside or work in areas that are exposed to higher levels of air pollution and thus exposure estimates for these populations may be underestimated. Furthermore, deprived populations may also experience poorer health due to other factors, such as smoking, less access to healthcare, and poorer dietary habits, which in turn can make these populations more susceptible to the detrimental effects of air pollution. The combination of all these factors further increase the risk of health problems in deprived populations (O'Neill et al., 2003).

Studies in Scotland do not tend to focus on disease incidence, but rather focus on either hospital admissions or mortality. The studies that utilised an ecological areal unit design observed substantial pollutant-health effects for NO_2 and PM_{10}, but at a greater spatial resolution than the spatial resolution used in this thesis (Huang et al., 2015; Lee, 2012; Lee et al., 2009; Lee & Mitchell, 2014; Lee et al., 2014). However, there have been studies conducted elsewhere that have primarily focused on the incidence of disease. Five studies have been identified, where only one study utilised an ecological areal unit design. The remaining four studies are cohort studies and mixed results were found. A meta-analysis conducted by Cesaroni et al. (2014) utilised 11 cohorts across Europe as part of the ESCAPE project, which investigated the long-term impact of $PM_{2.5}$ on the incidence of acute coronary events. The patients recruited for the study were free from any coronary events at the beginning of the study, and took part between 1997 and 2007. The overall pooled effect size for a $5\mu gm^{-3}$ increase in $PM_{2.5}$ concentrations was associated with a 13% increased risk of coronary events, with corresponding 95% confidence interval (0.98, 1.30). The authors concluded that long-term exposure to $PM_{2.5}$ was associated with incidence of coronary events, even though the levels of $PM_{2.5}$ observed were below the current European limit values. Another European cohort was devised in Rome between 1998 and 2000, which observed a borderline association (relative risk (RR) = 1.03, 95% CI = (1.00, 1.07)) between NO_2 concentrations and coronary events (Rosenlund et al., 2008). Weak relationships were also observed in a cohort study conducted in England by Atkinson et al. (2013), which investigated a number of cardiovascular diseases, but found only incidence of heart failure to be consistently associated with both PM_{10} and NO_2. Another cohort study, conducted by Miller et al. (2007), looked at the effect of $PM_{2.5}$ on the incidence of cardiovascular events in women in 36 US metropolitan areas. Again, the individuals did not have any cardiovascular disease prior to the study. The authors found that a $10\mu gm^{-3}$ increase in $PM_{2.5}$ was associated with a 24% increase in the risk of a cardiovascular event (hazard ratio (HR) = 1.24, 95% CI = (1.09, 1.41)). However, there was

an even stronger association when only cardiovascular deaths were considered (hazard ratio (HR) = 1.76, 95% CI = (1.25, 2.57)). The results from these cohort studies are broadly consistent, but none of them studied the incidence of cardio-respiratory disease. The only ecological study was conducted by Maheswaran et al. (2012) in Sheffield, England between 1995 and 2004. The study investigated the association between PM_{10} and NO_2 concentrations with incidence of ischemic and haemorrhagic stroke. The authors found no consistent associations between air pollution and stroke incidence, but did note the risks were slightly increased in the older age group (65-79 years). However, this health outcome is not the focus in the present analyses.

The chapter is organised as follows. Section 6.2 outlines the motivating study by discussing the disease, air pollutant and covariate data used in the statistical modelling. Section 6.3 briefly describes the ecological spatial model used to investigate the association between air pollution and ill health (a full discussion can be found in Chapter 5). Section 6.4 provides some descriptive and formal results from applying the spatial model to the health data. Finally, Section 6.5 provides a concluding discussion, and motivations for future research.

6.2 Motivating study

The methodology developed in this chapter is motivated by a new epidemiological study investigating whether long-term exposure to the air pollutant NO_2 has a detrimental impact on the incidence of cardio-respiratory disease. This extends the research conducted in Chapter 5 in terms of the study data, but focuses on incidence of cardio-respiratory disease.

The study area is West Central Scotland (discussed in Chapter 5 Section 5.2). Briefly, West Central Scotland is partitioned into $m = 2089$ non-overlapping areal units, known as data zones, that comprise 800 inhabitants on average. These data zones were constructed in such a way as to ensure homogeneity of the areal unit in terms of its social characteristics, and are presented in Figure 6.1. The remainder of this section describes the disease, air pollutant and covariate data to be used in the statistical modelling.

6.2.1 Disease data

This study seeks to investigate the incidence of cardio-respiratory disease (International Classification of Diseases, 10th Revision (ICD-10): I00-I99, J00-J99), by utilising hospital admission data and mortality records for the number of first events of cardio-respiratory disease between 2006 and 2012. These first events serve as a proxy measure of incidence, since they comprise either the first admission into hospi-

tal, where cardio-respiratory disease is the main cause, or cardio-respiratory deaths. Since these are strictly first events they exclude any patients with a hospital discharge for cardio-respiratory disease prior to the study period. Therefore, these data contain only the first record in the study period (commencing the year 2006 until 2012) - either a hospital admission or death - of all patients who had no known prior admission to hospital for cardio-respiratory disease. These data were obtained from National Services Scotland (NSS, https://nhsnss.org/), by submitting an application to the Privacy Advisory Committee (now known as the Public Benefit and Privacy Panel for Health and Social Care), which aims to protect personal individual level data when making it available for research. The hospital admission data are available from the SMR01 dataset, which is national dataset comprising all general/acute inpatient and day cases to hospital. The death records are available from National Records Scotland (https://www.nrscotland.gov.uk/), and were linked with the hospital admission data through the Information Services Division (http://www.isdscotland.org/). All analyses in this chapter were performed within the NSS National Safe Haven (http://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/#NSS-National-Safe-Haven), which is a secure environment where identifiable patient data are linked and accessed, while maintaining top level confidentiality.

The incidence data were available as counts of cardio-respiratory first events stratified by month and year of admission or death, sex (male or female), 5-year age groups (0-4, 5-9, ..., 90+), and data zone. In total, there were 161,752 cardio-respiratory first events between 2006 and 2012, equating to 23,100 first events per year on average, where 80,121 (49.55%) were male and 81,581 (50.45%) were female. However, due to the low numbers of first events occurring in any single strata (a mean count of less than one) the cardio-respiratory first events were aggregated over month, year, age group and sex in order to ensure there was enough variation in the response. This methodology is in line with Chapter 5 Section 5.2, which aggregated the cardio-respiratory deaths across the seven year study period. Table 6.1 shows the distribution of cardio-respiratory first events aggregated over month, age group and sex, but separately across the seven year period. This shows that the total number and distribution of first events are consistent across the study period. Further, it highlights a lack of temporal variation, which was also observed in the mortality data. While there is a considerably greater number of yearly first events, the distribution is more spread out across the data zones compared to when only the mortality data were considered. Further aggregation across the years results in greater variation compared to the mortality data (Table 5.1).

Fully-aggregated first events over all months, years, sex and age groups are denoted by Y_i for each data zone. Expected numbers, E_i, of cardio-respiratory first events were calculated based on the indirect standardisation method (see Chapter 2 Section 2.5

Table 6.1: *Summary statistics and total number of cardio-respiratory first events aggregated over age group and sex, but separately for each year. Summary statistics and the standardised incidece ratio (SIR) for the fully-aggregated first events over all age groups, sex, months and years are also displayed.*

Year	Min	25%	Median	Mean	75%	Max	Total
2006	0	0.740	0.969	1.003	1.242	2.962	23,205
2007	0	0.754	0.975	1.000	1.227	2.261	23,358
2008	0	0.738	0.976	1.003	1.236	3.272	23,856
2009	0	0.739	0.968	1.000	1.236	2.511	23,060
2010	0	0.743	0.972	1.003	1.236	2.910	22,314
2011	0	0.744	0.988	1.006	1.233	2.394	22,632
2012	0	0.759	0.984	1.009	1.223	3.267	23,277
2006-2012	0	58	74	77.41	93	270	161,702
SIR	0.289	0.840	0.992	0.999	1.147	2.075	

for a detailed description) using age- and sex-specific cardio-respiratory incidence rates for the whole of West Central Scotland. These expected numbers were calculated in order to take the varying population size and demographic structure of data zones into account. Specifically, $E_i = \sum_r N_{mr}\gamma_r$, where N_{nr} is the number of people in data zone n from age-sex strata r, while γ_r is the strata-specific disease rate for West Central Scotland. An exploratory measure of disease risk is the standardised incidence ratio (SIR), computed as $SIR_i = Y_i/E_i$, where an SIR of 1.2 corresponds to a 20% increase in the risk of disease compared to what is expected. Table 6.1 also showcases the distribution of the SIRs for the fully aggregated first events, which shows that, on average, (SIR = 0.999) what is observed in West Central Scotland is what is expected.

As discussed in Chapter 3 Section 3.3.2, the effect of air pollution on ill health has been shown to differ across the age spectrum, with the elderly population being more susceptible to its adverse effects (Fischer et al., 2003; Larrieu et al., 2007; O'Neill et al., 2004) due to having multiple co-morbidities. Therefore, to investigate whether the effect of air pollution is different at different ages in this specific analysis, the cardio-respiratory first events were aggregated across all months, years and sex to three separate age groups that represent three broad stages within a population: namely, younger population (0-19 years), working population (20-64 years), and older population ($\geqslant$ 65 years). The distribution across the three age groups is shown in Table 6.2, where a greater number of first events is observed for the working age group. For the younger age group, on average, there were 4898 first events per year, 10,311 per year on average for the working age group, and 8113 per year on average for the older age group.

The spatial distribution of disease risk is shown in Figure 6.1, which displays SIRs across West Central Scotland for the fully-aggregated first events, and the first events separately for the three age groups. All maps display similar spatial patterns, where

Table 6.2: *Summary statistics and total number of cardio-respiratory first events, stratified by three age groups: younger (0-19 years), working (20-64 years), and older (⩾ 65 years). Av. per year represents the average number of first events per year.*

Age (years)	Min	25%	Median	Mean	75%	Max	Av. per year	Total
0-19	0	8	12	13.750	17	140	4898	28,688
20-64	0	26	33	35.550	41	105	10,311	72,179
⩾ 65	0	17	26	29.100	37	163	8113	56,790

the data zones with the highest SIRs occur in areas with higher levels of deprivation. Furthermore, greater spatial variation is seen when the first events are separated according to age compared to the fully-aggregated SIRs which seem visually smoother.

6.2.2 Air pollutant data

The air pollutant utilised in this study is nitrogen dioxide (NO_2, measured in μgm^{-3}), in which concentrations are available at the yearly level between 2006 and 2012 inclusive. These are based on the statistical fusion model (denoted *Fusion*) developed in Chapter 4. This fusion model combines measured data obtained directly from air pollutant monitors and diffusion tubes located throughout the study region, and modelled concentrations from an atmospheric dispersion model, which predicts NO_2 concentrations on a regular 1km square grid. Predicting on a regular grid ensures complete spatial coverage of West Central Scotland, which is not possible when only considering directly measured data. However, modelled concentrations should not be considered on their own as they are estimated concentrations form a mathematical model, with no measure of uncertainty surrounding the estimates. Concentrations were averaged across the seven year period, and aggregated from the grid level to the data zone level. The spatial distribution of NO_2 concentrations is displayed in Figure 6.2, where the City of Glasgow has high levels of NO_2. The lower pollutant levels reflect the more rural parts of West Central Scotland, especially in the southern region. Concentrations of NO_2 range from 12 μgm^{-3} to 47μgm^{-3}, with a median concentration of 27 μgm^{-3}. As in the previous chapter, modelled concentrations (denoted *DEFRA*), are also used in addition to the fusion model concentrations so that comparisons can be made between the two analyses. Even though these data should not be considered on their own, they are also used here as a way of keeping in line with previous pollutant-health studies. When BMA was performed on the mortality data, both the *Fusion* and *DEFRA* concentrations contributed to the overall relative risk for cardio-respiratory mortality.

6.2.3 Deprivation data

As discussed in the Literature Review in Chapter 3 Section 3.3.4 and again in the previous chapter, socio-economic deprivation is an important factor to take into ac-

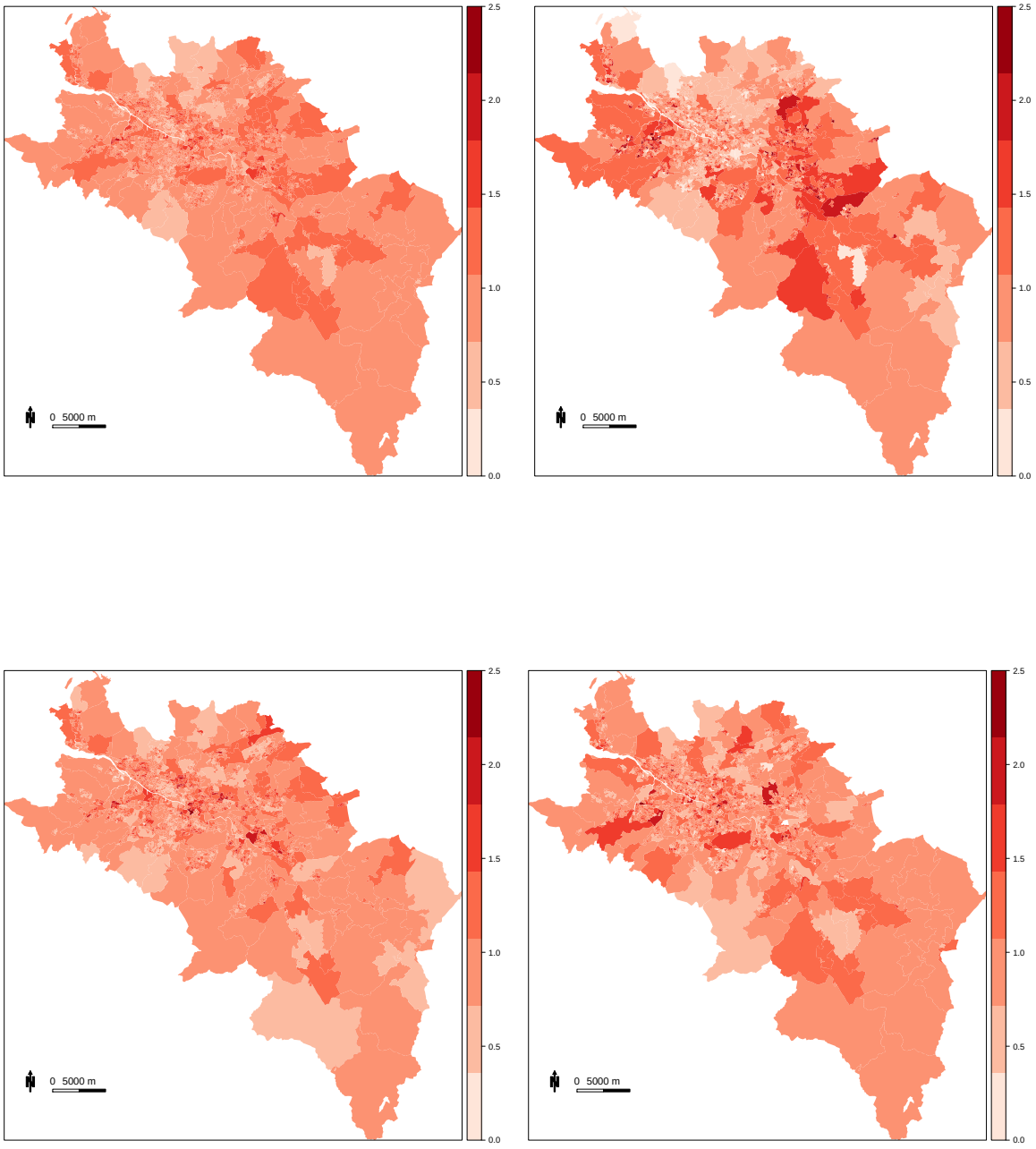

Figure 6.1: *Maps display the SIR for the cardio-respiratory first events, stratified by data zone (top left), then stratified by data zone and younger age group (0-29 years, top right), working age group (20-64 years, bottom left), and the older age group (⩾ 65 years, bottom right).*

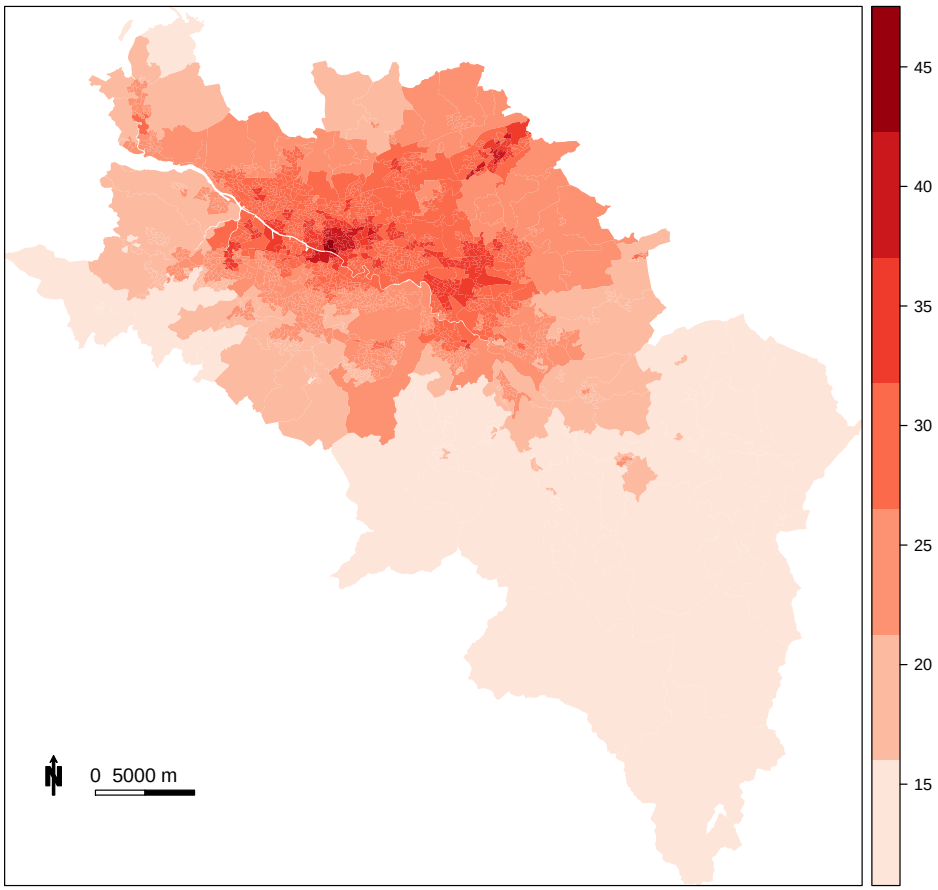

Figure 6.2: *Spatial map of the averaged 2006-2012 NO_2 concentrations from the statistical fusion model across West Central Scotland.*

count when investigating the relationship between air pollution and ill health. People living in more deprived neighbourhoods are more likely to experience worse health, on average, than people living in more affluent neighbourhoods, thus making them more vulnerable to the effects of air pollution (Laurent et al., 2007). Moreover, the inequality gap between the most and least deprived groups is increasing, with health improving faster in more affluent populations compared to more deprived populations (Leyland et al., 2007a). Health status is affected by, not only individual life choices, but also by contextual and ecological factors (Marmot, 2007). This chapter uses the 2009 Scottish Index for Multiple Deprivation (SIMD) to measure deprivation. This is an ecological measure of overall deprivation, which includes aspects of income, education and employment. However, the index also comprises a measure of health, which includes deaths that are part of the chosen outcome. Therefore, the index was re-weighted to exclude the health domain, where the methodology is discussed fully in Chapter 5 Section 5.2.3. The domains utilised here are income; employment; education, skills and training; housing; geographical access to services; crime; and the overall (minus health) domain. Details on these domains can be found in Chapter 5. The spatial distribution of the income domain is displayed in Figure 6.3, where the City of Glasgow contains the greatest number of data zones that are income deprived.

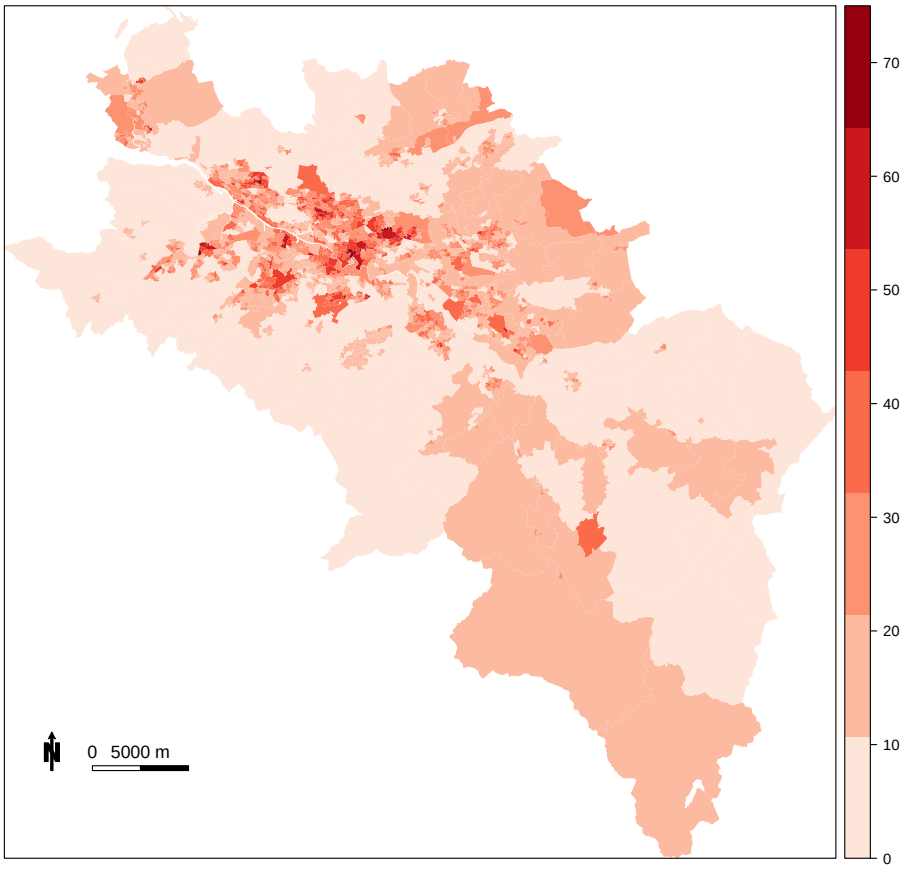

Figure 6.3: *Spatial map of the income domain across West Central Scotland. The income domain represents the percentage of each data zone's population who are in receipt of means-tested benefits, such as income support, income-based employment support allowance, and job seekers allowance.*

6.3 Statistical methods

The aim of this chapter is to investigate whether there is an association between the incidence of cardio-respiratory disease and exposure to NO_2 concentrations across West Central Scotland between 2006 and 2012 inclusive. In the previous chapter, the aims of the analyses were to, firstly, estimate the sensitivity of the NO_2-health effect by utilising two sets of NO_2 concentrations, utilising different deprivation indicators, and by changing the way in which to model residual spatial autocorrelation. Then, in order to determine a single effect estimate for NO_2 on mortality without having to choose one model based on a goodness-of-fit criterion, BMA was used to combine all models into an overall effect estimate that took model uncertainty into account. This method allows for the calculation of the model that contributes the most to the overall effect, thus highlighting which model can be deemed the most informative. The results showed clear sensitivity to model choice; however when all models were combined into a single effect estimate, only the income domain with the Leroux specification (using both sets of NO_2 concentrations) out of a possible 42 models determined the overall estimate.

Therefore, this analysis will only utilise the Leroux specification to model the resid-

ual spatial autocorrelation, while utilising both sets of NO_2 concentrations. Although the income domain was shown to be the only contributing socio-economic factor to cardio-respiratory mortality in the previous chapter, this chapter will consider all deprivation measures, since the outcome now includes hospital admissions in addition to the deaths, meaning that another deprivation measure may dominate instead of income. Furthermore, as this chapter also seeks to investigate whether the effect of NO_2 differs by different age groups, the deprivation measure which dominates may also change depending on the age group.

The remainder of this section will briefly describe the statistical model used to estimate the NO_2-health relationship, while also briefly describing BMA. Inference is performed within a Bayesian setting using McMC within the `R` software environment (R Core Team, 2015).

6.3.1 Spatial model

The vector of observed and expected numbers of cardio-respiratory first events (hospital admissions or deaths) is denoted by $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ and $\mathbf{E} = (E_1, \ldots, E_m)^\top$ respectively, where $m = 2089$ data zones in total. The vector of NO_2 concentrations (for either *Fusion* or *DEFRA*) is denoted by $\mathbf{x} = (x_1, \ldots, x_m)$ for all m data zones, while each measure of socio-economic deprivation is denoted by $\mathbf{u} = (u_1, \ldots, u_m)$. Thus, for the ith data zone, the vector of covariates is given by $\mathbf{z}_i^\top = (1, x_i, u_i)$, while the corresponding regression parameters are given by $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$, where β_1 is the intercept term, β_2 is the key parameter in this model, namely the effect of NO_2 on cardio-respiratory disease risk, and β_3 is the deprivation effect. A general Bayesian Poisson log-linear spatial model for these data is given by:

$$\begin{aligned} Y_i \mid E_i, R_i &\sim \text{Poisson}(E_i R_i) \quad \text{for } i = 1, \ldots, m, \\ \ln(R_i) &= \mathbf{z}_i^\top \boldsymbol{\beta} + \phi_i, \\ \boldsymbol{\beta} &\sim \text{N}(\mathbf{m}, \mathbf{V}), \\ \phi_i \mid \boldsymbol{\phi}_{-i} &\sim \text{N}\left(\frac{\rho \sum_{j=1}^m w_{ij} \phi_j}{\rho \sum_{j=1}^m w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^m w_{ij} + 1 - \rho} \right), \end{aligned} \tag{6.1}$$

where R_i is the risk of disease in data zone i. Again, the regression parameters, $\boldsymbol{\beta}$, are assigned a weakly informative multivariate Gaussian prior, with hyperparameters $(\mathbf{m}, \mathbf{V})$, typically with mean zero and a large diagonal variance matrix (such as $diag_{\boldsymbol{\beta}}(1000)$). The final term in the linear predictor is the vector of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$, which controls the residual spatial autocorrelation in the data after accounting for covariate effects.

The Leroux (Leroux et al., 1999) specification is used for $\phi_i \mid \boldsymbol{\phi}_{-i}$ here, due to its

flexibility in accounting for both strong and weak spatial autocorrelation structures. This model is specified by a set of m univariate full conditional distributions, $p(\phi_i \mid \boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_m)$. Spatial autocorrelation is imposed using a binary $m \times m$ neighbourhood matrix, $\mathbf{W}$, whose ijth element, $w_{ij} = 1$ if areas (i, j) share a common border, and $w_{ij} = 0$ otherwise. This specification asserts that neighbouring areas have random effects that are partially autocorrelated, otherwise the random effects are conditionally independent. Furthermore, ρ controls the level of spatial autocorrelation, with $\rho = 0$ corresponding to spatial independence with mean zero and constant variance, and $\rho = 1$ corresponding to strong spatial autocorrelation (and simplifying to the intrinsic CAR model given by equation (2.84) in Chapter 2 Section 2.4.2.1). Weakly informative hyperpriors are assigned for τ^2 and ρ; typically an inverse-gamma(a, b) distribution for τ^2, and a uniform distribution on the unit interval for ρ.

6.3.2 BMA

Bayesian model averaging (BMA) provides a framework for combining estimates for a specific quantity of interest from multiple Bayesian models into a single overall estimate that takes model uncertainty into account. This methodology will be described briefly here, but a full description can be found in Chapter 5 Section 5.3.5.

Numerous models are considered here due to the spatial model being applied separately for each of the seven deprivation measures and for both sets of NO_2 concentrations: the *Fusion* concentrations, and the *DEFRA* modelled concentrations. In total, there are $K = 14$ models, where $\hat{\beta}_{2_k}$ reflects the estimate for the NO_2-health relationship for the kth model. The models are denoted by $(M_1, \ldots, M_K)$, with corresponding model parameters given by the set $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. The true effect of NO_2 concentrations on the risk of cardio-respiratory disease is given by β_2, with its posterior distribution given by

$$p(\beta_2 \mid \mathbf{Y}) = \sum_{k=1}^{K} p(\beta_2 \mid M_k, \mathbf{Y}) p(M_k \mid \mathbf{Y}). \tag{6.2}$$

The posterior distribution of β_2 from model K is given by $p(\beta_2 \mid M_k, \mathbf{Y})$, and $p(M_k \mid \mathbf{Y})$ is the posterior probability of model M_k. In other words, the posterior distributions for the NO_2 effects are averaged across the $K = 14$ models, while being weighted by their posterior model probabilities, given by

$$p(M_k \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid M_k)}{\sum_{l=1}^{K} p(\mathbf{Y} \mid M_l)}. \tag{6.3}$$

Typically, if information is available regarding which model should be more influential to the overall effect size, then that can be incorporated into (6.3). However, no such information is available here, therefore, each model is assumed to have an equal contri-

bution to the overall effect size. Next, $p(\mathbf{Y} \mid M_k)$ is approximated by J McMC samples as

$$p(\mathbf{Y} \mid M_k) \approx \frac{1}{J} \sum_{j=1}^{J} p(\mathbf{Y} \mid \boldsymbol{\theta}_k^{(j)}, M_k) p(\boldsymbol{\theta}_k^{(j)} \mid M_k), \tag{6.4}$$

which gives the marginal probability of the data, $\mathbf{Y}$ given model M_k averaged over all parameters, where the superscript (j) denotes the jth McMC sample. Thereafter, the posterior mean and variance for the true relationship β_2 can be computed as

$$\mathbb{E}[\beta_2 \mid \mathbf{Y}] = \sum_{k=1}^{K} \hat{\beta}_{2_k} p(M_k \mid \mathbf{Y}), \tag{6.5}$$

for the mean, and

$$\text{Var}[\beta_2 \mid \mathbf{Y}] = \sum_{k=1}^{K} (\text{Var}[\beta_2 \mid M_k, \mathbf{Y}] + \hat{\beta}_{2_k}^2) p(M_k \mid \mathbf{Y}) - \mathbb{E}[\beta_2 \mid \mathbf{Y}]^2, \tag{6.6}$$

for the variance. Then, the uncertainty surrounding the estimate for the overall NO_2-health relationship is given by the approximate 95% credible interval (CI) as

$$95\% \text{ CI} = \mathbb{E}[\beta_2 \mid \mathbf{Y}] \pm 1.96 \sqrt{\text{Var}[\beta_2 \mid \mathbf{Y}]}. \tag{6.7}$$

6.4 Results

This section presents the results from investigating whether long-term exposure to NO_2 concentrations has a detrimental effect on the risk of cardio-respiratory disease in West Central Scotland between 2006 and 2012. Section 6.4.1 provides some descriptive analyses, while Section 6.4.2 describes the sets of results obtained from fitting the spatial model to both sets of NO_2 concentrations and all seven deprivation measures to the fully-aggregated (over month, year, age group and sex) first events. Section 6.4.3 presents the same set of aforementioned models, but applied to the aggregated (over month, year and sex) first events separately for three age groups: younger age group (0-19 years), working age group (20-64 years), and older age group ($\geqslant$ 65 years). Section 6.4.4 presents the overall estimate of the effect of NO_2 on cardio-respiratory first events using the BMA approach for the fully-aggregated first events, and the three age groups separately. Each model was based on running five parallel Markov chains for 120,000 iterations, which included a burn-in period of 20,000 iterations. Thinning was applied to the remaining samples in each chain in which every 10th sample was kept in order to reduce their autocorrelation, thus producing a final set of 50,000 posterior samples to be used to determine the association between NO_2 concentrations and cardio-respiratory disease.

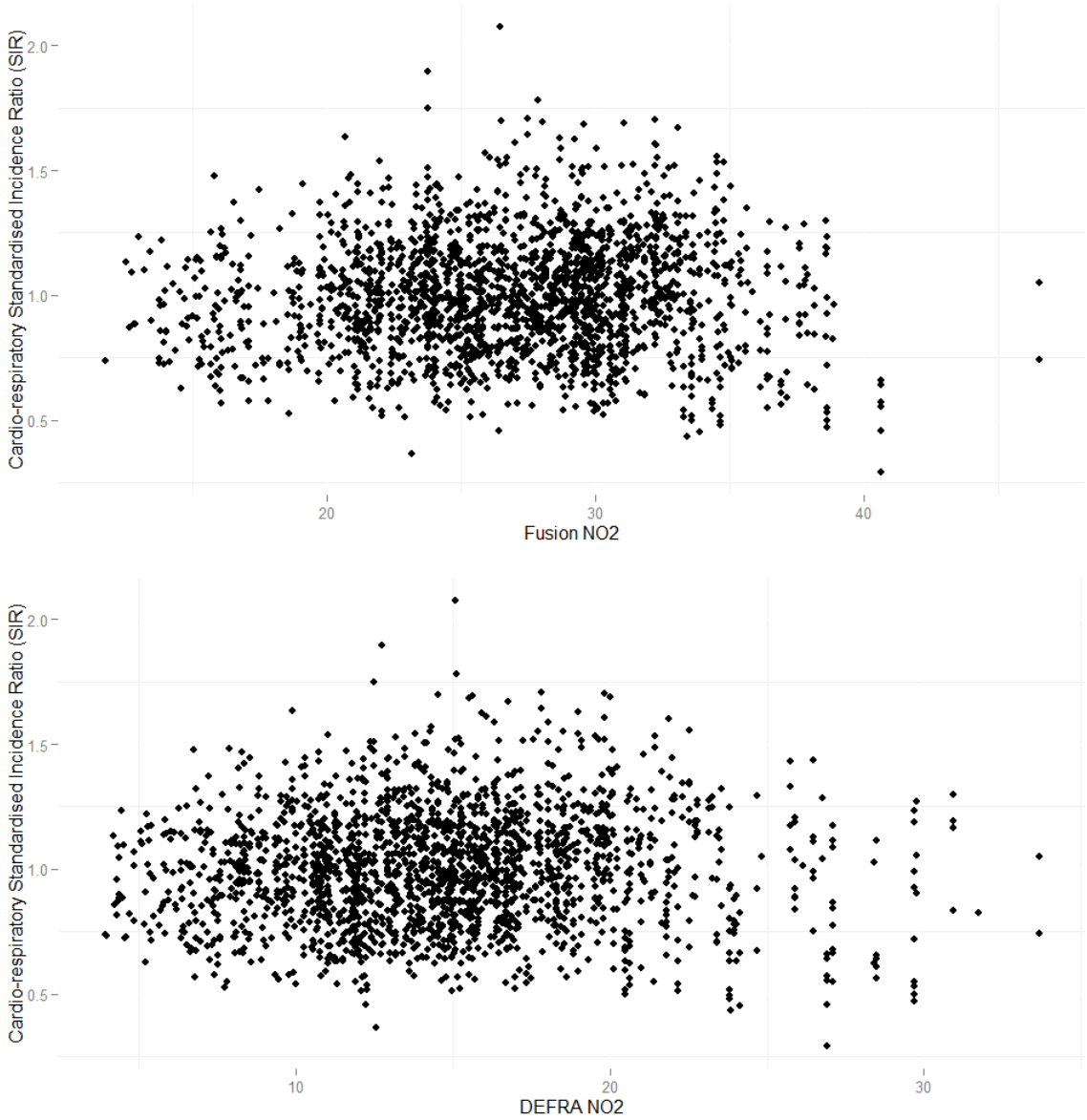

Figure 6.4: *Scatter plots displaying the relationship between the cardio-respiratory standardised incidence ratio (SIR) and both the Fusion NO_2 concentrations (μgm^{-3}) and the DEFRA NO_2 concentrations (μgm^{-3}) for the fully-aggregated first events.*

6.4.1 Descriptive results

This section explores relationships between variables, while performing simple Poisson regression in order to assess the strength of the residual spatial autocorrelation. In general, both the *Fusion* concentrations and the *DEFRA* concentrations exhibited a weak, but positive linear relationship with the cardio-respiratory SIR when investigating both the fully-aggregated first events, and the first events stratified by age group. This is shown in Figure 6.4 which displays scatter plots of the relationship between SIR and NO_2 concentrations for the fully-aggregated first events. Similar results were found for each of the three age groups separately.

When looking at the relationship between the SIR and the individual deprivation

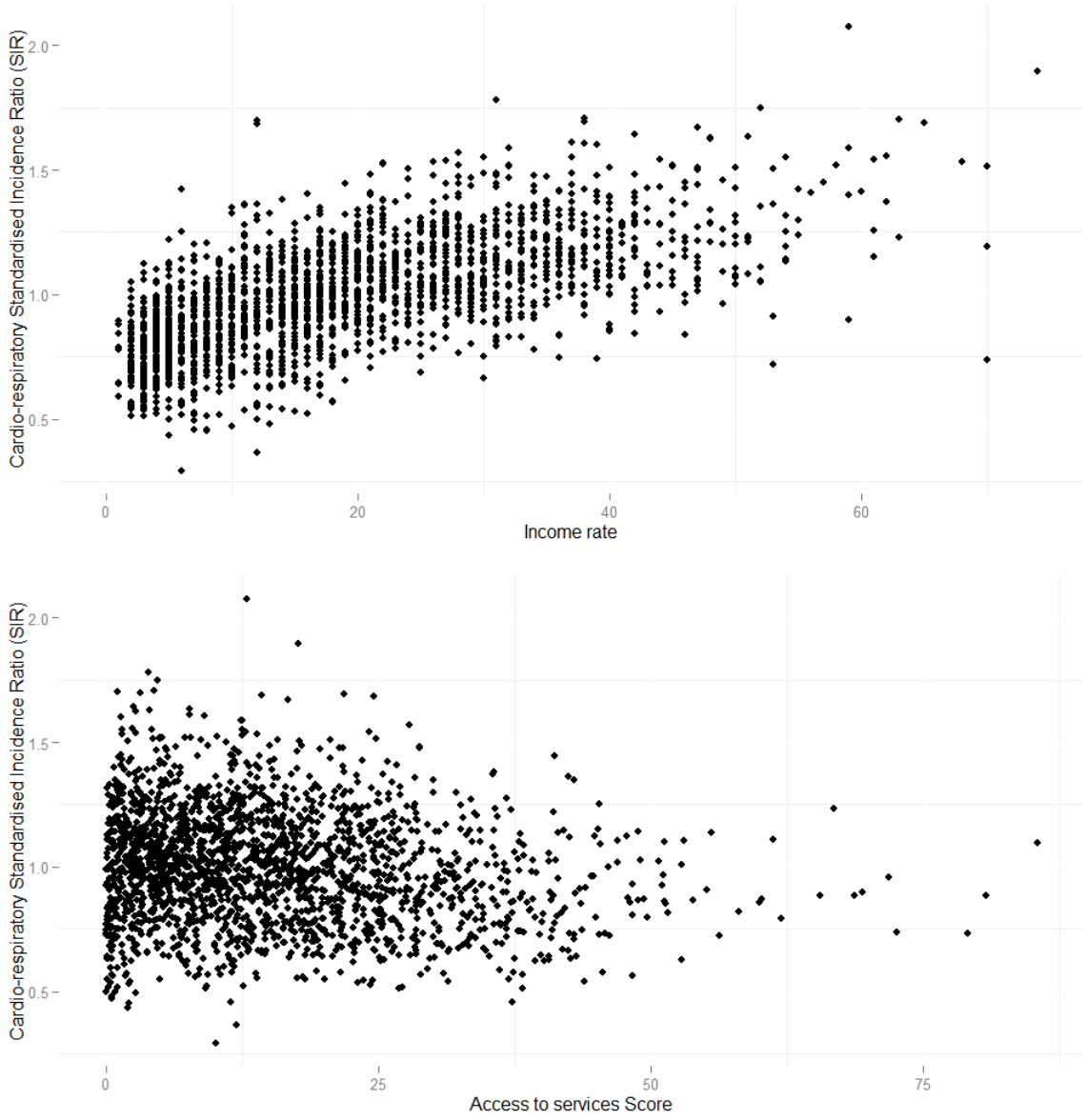

Figure 6.5: *Scatter plots displaying the relationship between the cardio-respiratory standardised incidence ratio (SIR) and two deprivation measures: income and access to services.*

measures, only access to services did not have a strong or positive relationship with SIR. This is displayed in the bottom panel of Figure 6.5, where it is clear that access to services has a slight negative, but linear relationship. The top panel is the relationship for the income domain, which shows a strong, positive and linear relationship with SIR. Both employment and education displayed similar relationships, with housing displaying a slightly weaker, positive relationship. Again, similar relationships were observed for all three age groups.

Initially, a simple Quasi-Poisson generalised linear model that did not include any spatial random effects was applied to the cardio-respiratory first events, with NO_2 (both *Fusion* concentrations and *DEFRA* concentrations) and each deprivation measure in turn as covariates. The results for the NO_2-health effects are given in Table 6.3. Inter-

estingly, none of the RRs are positive, suggesting that NO_2 has a protective effect on the risk of cardio-respiratory disease when fully aggregated over all years, sex and age groups. However, the overdispersion parameter ranges from 2.201 to 3.613, suggesting there is moderate overdispersion present with respect to the Poisson assumption of equal mean and variance. The dispersion is highest for the housing, access and crime domains, suggesting that these deprivation variables do not account for as much of the variation in cardio-respiratory first events compared to the remaining deprivation measures. Residuals from each model were then tested for residual spatial autocorrelation by calculating Moran's I statistic (Moran, 1950) given by equation (2.76) in Chapter 2 Section 2.4.2. Here, Moran's I statistic ranges from 0.203 to 0.338, and was statistically significant at the 5% level (according to the p-values) for all models, thus suggesting that spatial autocorrelation was present in the data. This is also shown in Figure 6.6, where it is clear that spatial autocorrelation remains. This map relates to the *Fusion* NO_2 concentrations with the income domain as the chosen deprivation measure, but similar results were observed for the other deprivation measures and when the *DEFRA* concentrations were used.

Table 6.3: *Quasi-Poisson generalised linear model results for the NO_2-health effect under each deprivation measure for the fully-aggregated data. Results show the relative risks (RR), 95% confidence intervals (CI), dispersion parameter, and Moran's I statistic for each model.*

NO_2	Deprivation	RR (95% CI)	Dispersion	Moran's I
Fusion	Income	0.971 (0.964, 0.978)	2.245	0.219
	Employment	0.973 (0.966, 0.980)	2.205	0.188
	Education	0.971 (0.964, 0.978)	2.404	0.161
	Housing	0.973 (0.963, 0.982)	3.394	0.320
	Access	0.987 (0.977, 0.997)	3.706	0.338
	Crime	0.992 (0.983, 1.001)	3.613	0.325
	SIMD	0.973 (0.973, 0.980)	2.278	0.209
DEFRA	Income	0.964 (0.957, 0.972)	2.205	0.210
	Employment	0.970 (0.962, 0.977)	2.201	0.185
	Education	0.969 (0.961, 0.977)	2.409	0.161
	Housing	0.961 (0.950, 0.971)	3.366	0.311
	Access	0.981 (0.970, 0.992)	3.702	0.336
	Crime	0.991 (0.981, 1.000)	3.613	0.324
	SIMD	0.968 (0.961, 0.976)	2.270	0.203

Similar results were observed for the younger and working age group. For the younger age group, RRs ranged from 0.917 to 0.948, with dispersion (1.980 - 2.265) and Moran's I (0.313 - 0.374) values indicative of residual spatial autocorrelation. The working age group had RRs ranging from 0.934 to 0.989, which suggests a weak association between NO_2 concentrations and cardio-respiratory disease in this age group. In addition, the dispersion (1.517 - 2.935) and Moran's I (0.142 - 0.341) values reduced, but were still statistically significant and suggested residual spatial autocorrelation was

present. However, in the older age group, RRs were positive, suggesting that exposure to NO_2 concentrations has a detrimental effect on cardio-respiratory disease. The RRs ranged from 1.002 to 1.036, with six out of 14 models containing the null risk of one in their corresponding confidence intervals. The dispersion (0.914 - 2.089) and Moran's I (0.051 - 0.096) values reduced further, while still suggesting spatial autocorrelation is present.

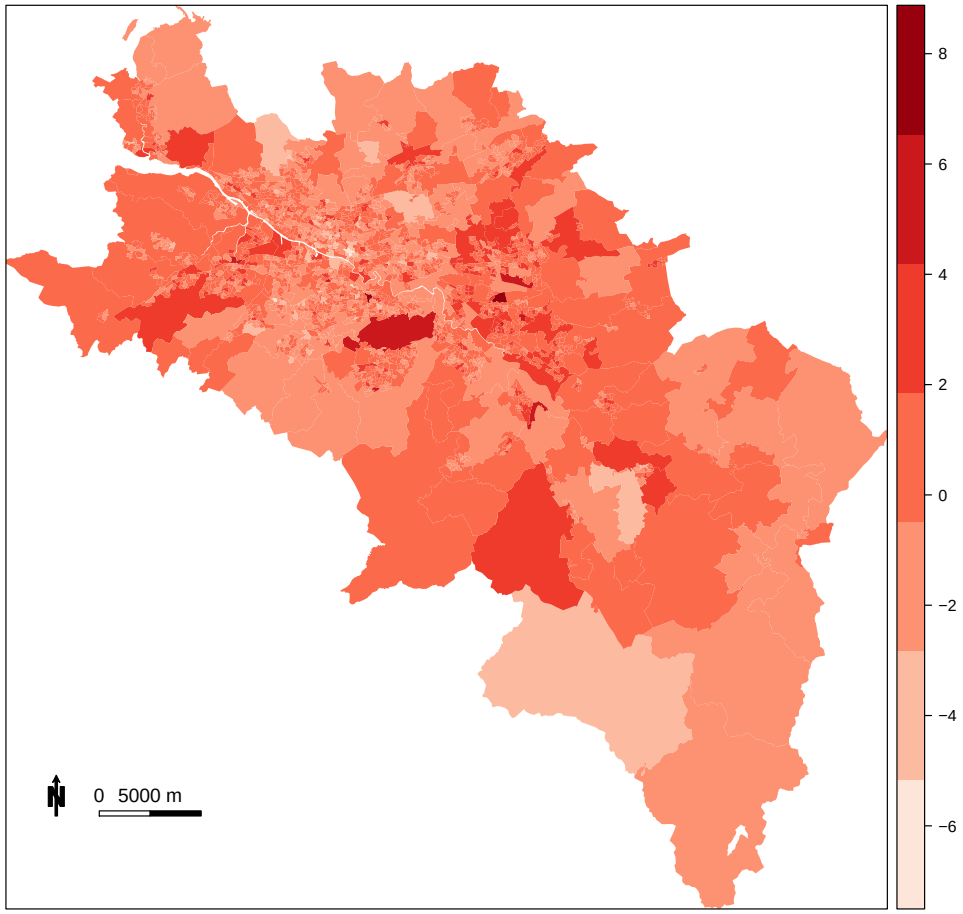

Figure 6.6: *Spatial map of residuals from a quasi-Poisson model, with Fusion NO_2 concentrations and the income domain as covariates on the fully-aggregated first events.*

6.4.2 Spatial model on fully aggregated first events

This section investigates the sensitivity of the NO_2-health effect according to how NO_2 is estimated, and the deprivation measure included as a covariate in the statistical

model for the fully-aggregated cardio-respiratory first events. Combination of these two factors results in a total of 14 models. Bayesian spatial models are used here, which include a set of spatial random effects to take into account any residual spatial autocorrelation in the data after the covariates have been accounted for.

As aforementioned, the Moran's I statistic suggested substantial spatial autocorrelation in the residuals when no random effects were included, and this was reflected in the spatial autocorrelation parameter, ρ, as all values were close to one implying strong spatial autocorrelation. The posterior median and 95% credible intervals for each combination of NO_2 and deprivation measure are given in Table 6.4, where the NO_2-health effects are presented on the relative risk scale for a 5 μgm^{-3} increase in NO_2 concentrations, since this reflects a realistic change in exposure. Overall, results suggest that NO_2 has a negative association with cardio-respiratory first events since the RRs are below the null risk of one, ranging from 0.977 to 0.996. However, out of the 14 models only two are substantial at the 95% level, since their corresponding credible intervals are wholly below the null risk of one. These two models relate to the education domain, which displays a 2.3% decreased risk in cardio-respiratory first events when the *Fusion* concentrations are used, and a 2.1% decreased risk when the *DEFRA* concentrations are used. Both of these models only show a small decreased risk, and are also only borderline substantial, since the upper limit of the credible interval is close to one. Meanwhile, all remaining models show no evidence of a substantial relationship between NO_2 concentrations and the risk of cardio-respiratory first events, while being adjusted for various indicators of deprivation. Furthermore, the range in estimated RRs is small, with a difference of only 1.9%, suggesting that the estimated relationship between NO_2 concentrations and the risk of cardio-respiratory disease is robust to the choice of NO_2 concentrations and indicator of deprivation. The only pattern of note is that the estimated effect sizes are slightly attenuated when the *Fusion* concentrations are used compared to the *DEFRA* concentrations.

Even though education was the only domain that resulted in a small, but substantial effect for NO_2 on the risk of cardio-respiratory disease, the statistical model including the income domain had the best fit to the data compared to the remaining models in terms of the deviance information criterion (DIC), since it had the lowest value. This finding is in line with what was observed in the previous chapter, when cardio-respiratory mortality was the chosen outcome. However, the observed association between NO_2 concentrations and cardio-respiratory mortality in the previous chapter was positive, unlike here, where the association was found to be negative. However, both analyses were in agreement that the statistical model including the income domain had the best fit to the data, in terms of DIC.

Table 6.4: *Bayesian Poisson model results for the NO_2-health effect under each deprivation measure for the fully-aggregated data. Results show the relative risks (RR), 95% credible intervals (CI), spatial correlation parameter ρ and the deviance information criterion (DIC) for each model. RRs are presented for a $5\mu gm^{-3}$ increase in NO_2 concentrations.SIMD represents the entire deprivation index re-weighted without the health domain. The results in bold are substantial at the 5% level.*

NO_2	Deprivation	RR (95% CI)	ρ (95% CI)	DIC
Fusion	Income	0.989 (0.975, 1.005)	0.926 (0.932, 0.979)	15906
	Employment	0.988 (0.974, 1.002)	0.848 (0.716, 0.941)	15917
	Education	**0.977 (0.965, 0.990)**	0.603 (0.440, 0.764)	16070
	Housing	0.990 (0.971, 1.011)	0.923 (0.845, 0.977)	16196
	Access	0.981 (0.961, 1.001)	0.985 (0.811, 0.960)	16247
	Crime	0.989 (0.971, 1.008)	0.875 (0.781, 0.949)	16230
	SIMD	0.992 (0.976, 1.008)	0.885 (0.773, 0.961)	15935
DEFRA	Income	0.992 (0.976, 1.010)	0.928 (0.834, 0.980)	15906
	Employment	0.990 (0.975, 1.007)	0.982 (0.710, 0.946)	15916
	Education	**0.979 (0.966, 0.993)**	0.604 (0.440, 0.775)	16072
	Housing	0.994 (0.973, 1.015)	0.925 (0.847, 0.978)	16199
	Access	0.986 (0.966, 1.009)	0.892 (0.807, 0.958)	16249
	Crime	0.996 (0.976, 1.017)	0.875 (0.781, 0.950)	16224
	SIMD	0.995 (0.977, 1.012)	0.892 (0.779, 0.965)	15937

6.4.3 Spatial model on first events stratified by three age groups

This section explores how the association between NO_2 concentrations and the risk of cardio-respiratory disease changes across the age spectrum. Data are stratified into three age groups: younger (0-19 years), working (20-64 years), and older ($\geqslant$ 64 years), where the sensitivity of the association will be investigated according to the choice of NO_2 concentrations (either *Fusion* or *DEFRA*), and deprivation indicator. The combination of these factors results in a total of 42 models, where the statistical model of choice is the Bayesian spatial model that incorporates spatial random effects.

The Moran's I statistics for the quasi-Poisson models ranged from 0.313 to 0.374 for the younger age group; 0.341 to 0.142 for the working age group; and 0.051 to 0.096 for the older age group. All statistics were statistically significant; however, the level of residual spatial autocorrelation reduces as the age of the population increases. A similar pattern is observed in the Bayesian spatial models for the ρ parameter, which ranged from 0.993 to 0.996 for the younger age group; 0.887 to 0.976 for the working age group; and 0.007 to 0.155 for the older age group. These results for the younger and working age group are in line with what was observed for the fully-aggregated data, whereas the Moran's I results for the older age group are in line with what was observed when using mortality data only (statistic of 0.036).

The results for these models are displayed in Table 6.5, where the RRs and 95% credible intervals are shown for all three age groups. The RRs are consistent within the specified age groups, but vary considerably across the age groups. For the younger age group, the results differ from the quasi-Poisson results in terms of the relationship being estimated to be positive instead of negative. Therefore, increasing NO_2 concentrations is associated with a small, but positive increase in the risk of cardio-respiratory disease; however, the evidence suggests the association is not substantial, since the 95% credible intervals contain the null risk of one. The working age group displays a negative association with the risk of cardio-respiratory disease, with all posterior medians being less than one; however, only two out of the 14 models are borderline substantial. Therefore, there is little evidence in the working age group of a relationship between NO_2 concentrations and the risk of cardio-respiratory disease. For the older age group, the RRs switch to being positive, which is in line with the mortality analyses in the previous chapter, and implies that there is a small, but positive relationship between NO_2 concentrations and the risk of cardio-respiratory disease. Eight out of the 14 models have substantial relationships, with the crime domain having the strongest relationship. However, in terms of DIC, the income domain still has the lowest value, and is therefore considered as the best fit to the data. Conversely, the DIC was lowest for the employment domain in the working age group, but was lowest for income in the younger age group. These results suggest that it is important to consider the pollution-health relationship at different ages in order to understand where the relationship lies, which in turn can help policy makers target the people that are most affected by the detrimental effects of air pollution.

In addition to the RRs displayed for the NO_2-health effects, Table 6.6 displays the corresponding deprivation-health effect separately for the three age groups. The deprivation-health effects are extremely similar across the two measures of NO_2 concentrations. The RRs are also similar across the three age groups for each deprivation indicator; however, the effect sizes are slightly higher for the working age group, and are slightly attenuated for the older age group. The access to services domain was the only deprivation indicator with a negative relationship with the risk of cardio-respiratory disease; however, this relationship is not strong since the upper limit of the 95% credible intervals is close to the null risk of one. These results are in line with the scatter plots displayed in Figure 6.5, which shows a slightly decreasing albeit weak relationship between SIR and the access to services score. This also highlights that the access to services domain exhibits an independent spatial pattern compared to the other indicators, which was also indicated in the previous chapter when mortality was the outcome. Six out of the seven deprivation indicators have narrow credible intervals, whereas the crime domain has credible intervals that are extremely wide. This highlights greater uncertainty for this domain.

Table 6.5: *Relative risks (RR) and 95% credible intervals for the NO_2-health effect under each deprivation measure for the younger (0-19 years), working (20-64 years) and older ($\geqslant$ 65 years) age groups. RRs are presented for a 5μgm^{-3} increase in NO_2 concentrations. SIMD represents the entire deprivation index re-weighted without the health domain. The results in bold are substantial at the 5% level.*

NO_2	Deprivation	RR (95% CI)		
		Younger	Working	Older
Fusion	Income	1.019 (0.988, 1.051)	0.988 (0.970, 1.008)	1.008 (0.996, 1.019)
	Employment	1.012 (0.990, 1.055)	0.988 (0.969, 1.005)	1.011 (1.000, 1.023)
	Education	1.011 (0.981, 1.043)	**0.979 (0.962, 0.999)**	**1.013 (1.002, 1.025)**
	Housing	1.022 (0.988, 1.055)	0.996 (0.972, 1.021)	1.002 (0.989, 1.014)
	Access	1.012 (0.974, 1.048)	0.981 (0.956, 1.007)	1.009 (0.996, 1.022)
	Crime	1.018 (0.985, 1.053)	0.996 (0.970, 1.020)	**1.018 (1.002, 1.032)**
	SIMD	1.026 (0.996, 1.058)	0.993 (0.975, 1.011)	1.011 (1.000, 1.023)
DEFRA	Income	1.023 (0.990, 1.057)	0.985 (0.966, 1.004)	**1.018 (1.009, 1.011)**
	Employment	1.026 (0.993, 1.058)	0.985 (0.976, 1.004)	**1.023 (1.010, 1.035)**
	Education	1.017 (0.984, 1.051)	**0.974 (0.955, 0.995)**	**1.026 (1.013, 1.039)**
	Housing	1.023 (0.989, 1.055)	0.991 (0.965, 1.018)	1.007 (0.993, 1.021)
	Access	1.018 (0.981, 1.055)	0.981 (0.954, 1.008)	**1.019 (1.005, 1.035)**
	Crime	1.025 (0.990, 1.065)	0.996 (0.969, 1.022)	**1.031 (1.016, 1.047)**
	SIMD	1.023 (0.994, 1.061)	0.989 (0.970, 1.008)	**1.022 (1.010, 1.035)**

6.4.4 BMA

Results from the previous section demonstrated that the effect of NO_2 on ill health differed by age group and was sensitive to the choice of deprivation indicator. When BMA was performed on the fully-aggregated first events, the overall estimated relative risk was 0.991, together with an associated 95% credible interval of (0.975, 1.008). This small, but the negative effect indicates that for a 5μgm^{-3} increase in NO_2 concentrations, the cardio-respiratory first events decrease by an estimated 0.9%. However, this overall result is not substantial since the credible interval encompasses the null risk of one, indicating that there is no evidence of a relationship between NO_2 and cardio-respiratory ill health. These results are in line with what was observed for the cardio-respiratory deaths, where the estimated RR was 1.011 (0.993, 1.029). Conversely, a positive relationship was observed here; however, the relationship was not substantial, since the credible interval contained the null risk of one. Again, income was the only deprivation indicator that contributed to the overall effect size in which the *Fusion* model contributed 45.44% and the *DEFRA* model contributed 54.56%. These results are in line with the mortality study, which observed a greater influence on the overall effect size from the *DEFRA* (67.83%) model compared to the *Fusion* (32.17%) model. However, in this case, the difference in influence between the two models is small.

BMA was performed separately for the three age groups, where for the younger age group, the estimated overall RR and 95% credible interval was 1.013 (0.982, 1.045);

Table 6.6: *Relative risks (RR) and 95% credible intervals for each deprivation measure for the younger (0-19 years), working (20-64 years) and older ($\geqslant$ 65 years) age groups. RRs are presented for a one standard deviation increase in deprivation. SIMD represents the entire deprivation index re-weighted without the health domain.*

NO_2	Deprivation	RR (95% CI)		
		Younger	Working	Older
Fusion	Income	1.010 (1.009, 1.011)	1.015 (1.014, 1.016)	1.008 (1.007, 1.009)
	Employment	1.013 (1.011, 1.015)	1.021 (1.020, 1.022)	1.010 (1.009, 1.011)
	Education	1.141 (1.121, 1.161)	1.203 (1.189, 1.218)	1.078 (1.066, 1.091)
	Housing	1.006 (1.005, 1.007)	1.007 (1.006, 1.008)	1.004 (1.004, 1.005)
	Access	0.997 (0.996, 0.999)	0.995 (0.994, 0.996)	0.995 (0.994, 0.997)
	Crime	1.828 (1.420, 2.320)	1.982 (1.677, 2.323)	1.992 (1.656, 2.392)
	SIMD	1.006 (1.005, 1.007)	1.010 (1.010, 1.011)	1.005 (1.004, 1.005)
DEFRA	Income	1.010 (1.009, 1.011)	1.015 (1.014, 1.016)	1.007 (1.007, 1.008)
	Employment	1.013 (1.011 1.014)	1.021 (1.020, 1.022)	1.010 (1.009, 1.011)
	Education	1.141 (1.122, 1.160)	1.203 (1.189, 1.217)	1.076 (1.064, 1.088)
	Housing	1.006 (1.004, 1.007)	1.007 (1.006, 1.008)	1.004 (1.004, 1.005)
	Access	0.997 (0.996, 0.999)	0.995 (0.994, 0.996)	0.996 (0.995, 0.997)
	Crime	1.820 (1.429, 2.300)	1.982 (1.677, 2.336)	1.937 (1.607, 2.327)
	SIMD	1.006 (1.005, 1.007)	1.010 (1.010, 1.011)	1.005 (1.004, 1.005)

0.984 (0.966, 1.003) for the working age group; and 1.008 (0.996, 1.019) for the older age group. For all age groups, the estimated RRs were small, but not substantial at the 95% level indicating no evidence of an association between NO_2 and cardio-respiratory first events even at different age groups. Furthermore, the deprivation indicator with the most influence changes across the three age groups. The education domain contributed to the overall effect size in the younger age group (21.17% for *Fusion* and 78.83% for *DEFRA*), while the employment domain had the most influence for the working age group (92.38% for *Fusion* and 7.62% for *DEFRA*). Similarly to the fully-aggregated BMA, the income domain had the most influence in the older age group (0.01% for *Fusion* and 99.91% for *DEFRA*). These results show that out of the numerous deprivation indicators, only one contributes to the overall effect size, which was also observed in the previous chapter.

6.5 Discussion

This chapter sought to investigate the association between NO_2 concentrations and the incidence of cardio-respiratory disease in West Central Scotland between 2006 and 2012 inclusive. Incidence was defined as the number of first events of cardio-respiratory disease in terms of death or the first admission into hospital, with cardio-respiratory disease as the main cause. Incidence was the focus of this chapter, since it studies the number of new cases of a disease in a population in order to understand the risk of developing cardio-respiratory disease as a result of air pollution. Similarly to the previous chapter on cardio-respiratory mortality, this chapter investigated the sensitivity of the NO_2-health relationship according to how NO_2 was estimated, the choice of

deprivation indicator, and whether the observed relationship was different at different age groups.

The main finding from this chapter is that the way in which NO_2 concentrations were estimated and which deprivation indicator was included did not have a major impact on the estimated relationship between NO_2 and cardio-respiratory disease. This is in contrast to the previous chapter when mortality was considered, which saw these two factors having a major impact on the resulting NO_2-health effects. BMA was utilised as a way of combining information from all models into a single, overall estimate, rather than choosing one model based on a goodness-of-fit measure and ignoring the uncertainty from having multiple estimated effect sizes. The final estimated effect size for the fully-aggregated first events shows that a $5\mu gm^{-3}$ increase in NO_2 concentrations is associated with 0.9% lower cardio-respiratory first events in West Central Scotland between 2006 and 2012. However, this effect is not substantial at the 5% level, since the resulting 95% credible interval contains the null risk of 1. Again, this could be due to the NO_2 concentrations being too low to show a strong impact, with greater variation being paramount to observe any substantial health impacts. Furthermore, this result is not in line with what was observed when mortality was considered, which saw a small, but positive effect of NO_2 concentrations on cardio-respiratory mortality. The correlation between mortality and first events is 0.713, which indicates a moderate to high correlation between the two variables. Even though there is a relatively high correlation between the two variables, slightly opposing effect sizes are still observed. This suggests that the hospital admission and mortality data behave differently to each other in this particular setting. However, in both analyses, there was no evidence to suggest a substantial association between NO_2 concentrations and cardio-respiratory disease. Huang et al. (2015) conducted a spatio-temporal study across the whole of Scotland, where the outcome of the study focused on all respiratory hospital admissions, and not just the first admission into hospital. This study found a small negative relationship when *Fusion* NO_2 concentrations were used (the authors developed their own fusion model), but found a small positive relationship when *DEFRA* concentrations were used. Together with the analyses conducted in this thesis, there is clear inconsistency in the results, which emphasises the need for further study, in terms of separating out the effects of respiratory and cardiovascular disease, as the two diseases could behave in different ways. Unfortunately, this was not possible in this thesis due to time constraints, but would inherently be important in future research. Furthermore, the difference between hospital admissions and mortality could be studied further in order to tease out where the effect of air pollution is greatest.

In contrast to the fully-aggregated first events results discussed above, the NO_2-health effect was not consistent across the three age groups. In the younger age group (0-19 years), which includes children and young adults, there was a small, but positive

effect on the risk of cardio-respiratory first events, while in the working age group, the effect was negative , which is in line with the fully-aggregated results. Conversely, the older age group ($\geqslant$ 65 years) found a small, but positive effect, which is in line with the results from using mortality data as the study outcome. The difference in effect sizes between the age groups could be due the nature of the three different age groups. For example, individuals in the younger age group may have greater admissions into hospital with acute respiratory illness, such as asthma, rather than having illnesses that cause death. Individuals within the working age group may have a mix of cardiovascular and respiratory diseases and therefore greater numbers of hospital admissions; however, cardiovascular disease tends to be more prominent by the late 40s. The older age group may be likely to include the greatest numbers of deaths, with fewer numbers of admissions (which would most likely relate to cases of pneumonia, since it is an acute condition, Gittins et al., 2013), and in general, tend to suffer from more cardiovascular diseases. Moreover, there is a great deal of exposure misclassification with regards to the younger and working age group, but not so much for the older age group. Exposure is generally more accurate for older people, since the majority of their time is spent at home and they tend to stay within the same data zone. However, the younger age group will see individuals spending most of their time at school, which may not be within the same data zone as their home, especially for individuals that reside within the city as the data zones become smaller with increasing population density. Likewise, for the working age group, exposure misclassification is prominent since individuals are more likely to travel from the home to the workplace, which is often not in the same data zone. It is likely that this type of measurement error biases the relative risks towards the null risk of one, indicating no substantial association (Armstrong, 1998).

Previous studies have observed stronger pollutant-health effects in the elderly population (Fischer et al., 2003; Larrieu et al., 2007; O'Neill et al., 2003); however, the results observed here, are in line with Maheswaran et al. (2012), who found no association between air pollutants and ischemic stroke incidence in London. The authors noted that while there was no substantial association, the effects were slightly stronger in the older age group (65-79 years). In addition, a small area ecological study by Lawson et al. (2012), observed a small negative association between $PM_{2.5}$ and asthma incidence aggregated across all ages, which is in line with the fully-aggregated analysis presented here. Lee & Shaddick (2010) investigated the relationship between respiratory mortality and a number of air pollutants in the Greater London region between 2003 and 2005. The age group under study was the same as the older age group defined here, and the authors found a small positive effect (RR of 1.013 (0.998, 1.027)) of NO_2 concentrations on respiratory mortality. This effect size is dissimilar to what was observed in this study, but is similar in the sense that no substantial association was observed.

With regards to the relationship between air pollution and children's health, Beatty & Shimshack (2014) have noted that the relationship has been understudied, with the main focus being on adult and older age groups. The authors note that children could be vulnerable to air pollution, having smaller than average lung size as they develop into adulthood, as well as having a more active lifestyle (Committee on Environmenta Health, 2004). However, in this chapter, no substantial association was observed.

It has been shown that exposure to air pollution tends to have a greater effect on more deprived groups, with socially and economically deprived people more likely to live in more polluted areas, whether on purpose or accidentally (Liverani et al., 2016; O'Neill et al., 2003). This chapter observed consistent NO_2-health effects when different deprivation indicators were considered. However, this was in contrast to what was observed in the previous chapter when mortality was the study outcome, which found that the resulting NO_2-health effect was highly variable depending on the deprivation indicator used. In general, individuals from more deprived backgrounds are more likely to be admitted to hospital. This could, in part, be due to to individuals from less deprived backgrounds having greater knowledge of when to seek support and seeking support earlier for poor health. Therefore, it is possible that referral bias in present in these analyses.

It has been established that low levels of income and education are associated with higher risks of mortality and morbidity (Fernández-Somoano et al., 2013; O'Neill et al., 2003). For both the fully-aggregated first events and the fully-aggregated mortality data, income was the deprivation indicator with the greatest influence on the overall NO_2-health effect, which is in line with the literature (Fernández-Somoano et al., 2013; O'Neill et al., 2003). When the first events were stratified according to the three age groups, the education and employment indicators had the greatest influence for the younger and working age group, while income had the greatest contribution in the older age group, which is in line with the fully-aggregated first events and mortality. While results are different for two out of the three age groups, it is important to note that the correlation between education, employment and income was high, ranging from 0.833 to 0.946, which suggests that these three variables are essentially providing the same information, and thus, indicating robustness in results between the three age groups. These results also suggest that other studies focusing on income, employment or education as their main measure of socio-economic deprivation, are utilising the best possible variables when accounting for deprivation in their analyses.

Similarly to the previous chapter focussing on cardio-respiratory mortality, the estimated effect sizes were slightly attenuated when the *Fusion* NO_2 concentrations were used compared to the *DEFRA* concentrations. Both sets of pollutant concentrations influenced the overall effect size; however, how much each influenced the overall effect

size varied across the age groups, with the *DEFRA* concentrations having the majority of the greatest influences. This is in line with the previous chapter, which also found that the *DEFRA* concentrations had the greatest influence. These results are broadly consistent with previous studies, such as Huang et al. (2015), indicating that studies which use modelled concentrations, such as *DEFRA*, will obtain similar results compared to taking the proper steps to account for uncertainty in the modelled pollution estimates.

There are a few limitations to the analyses conducted in this chapter. Firstly, the cardio-respiratory first events were aggregated over a 7 year period to ensure that there was enough variation in the response. The aggregated first events were then stratified according to three broad age groups; however, it would be interesting to utilise narrower age groups in order to get a more detailed understanding of where the effect of air pollution is greatest. This would help aid policy makers in tackling the burden of detrimental air quality. However, reducing the age groups further will hinder the variation in the response, therefore either more data are required or different outcomes should be studied. Again, only a purely spatial study could be conducted due to the small number of first events in each data zone, therefore, aggregating to a higher spatial resolution may improve the power of the study. Secondly, cardio-respiratory disease is a combination of two main disease types, but it would be interesting to study these two diseases separately. It was not possible due to time constraints with the Safe Haven and data controllers; however, high level aggregation would have to take place to ensure enough variation in the response for detecting any potential substantial effects.

Suggestions for future work include the study of vulnerability, by following up a cohort of people that have already been admitted to hospital for cardio-respiratory disease and observing whether they have further admissions or not. This would allow the study of air pollution on a population that is already considered vulnerable due to having co-morbidities, which is more representative of the population as a whole.

Chapter 7

Conclusion

7.1 Introduction

There has been much research surrounding the detrimental effects of air pollution on the human population, where long-term exposure has been linked with increased risks of both cardio-respiratory mortality and morbidity. Exposure to air pollution has been a global concern for the past eighty years, and initially came light due to extreme air pollution episodes, including the London Smog in December 1952 (Ministry of Public Health, 1954). These short-term air pollution episodes, usually lasting over a period of two to three days, saw a rise in the numbers of people with respiratory distress and increased the numbers of premature deaths. These air pollution episodes and the research surrounding them have helped develop legislation, such as the Clean Air Act (1993) in the UK, to improve air quality by setting target levels that have to be met. Areas which fail to achieve these targets are investigated, and strategies initiated to try and mitigate potential hazardous effects.

Generally, research into the effects of air pollution on ill health utilises either, time series studies when investigating the effects of short-term exposure, or cohort studies at the individual level when investigating the longer-term impact of air pollution on ill health. Although cohort studies are extremely important research tools, allowing for inferences to be made at the individual level, as well as helping establish causal relationships, they can be time-consuming due to the need to follow-up patients over a number of years. More recently, spatial ecological studies are being used, where a population level association is estimated between air pollutant levels and ill health. Data for these studies are more widely available and analyses are quick to implement as there is no long follow-up period. While these studies are performed at the group level, meaning cause and effect cannot be directly established, they are still valuable in terms of making a contribution to the growing body of evidence, which demonstrates that exposure to air pollution, even at low levels, is detrimental to health.

The majority of studies into the effects of air pollution on ill health in the United

Kingdom have been conducted in England, with a particular focus on London given its status of being the largest city in the UK in terms of population number and density. Relatively few studies have been conducted in Scotland. Scotland is predominantly a rural country, with the majority of the population residing within the central belt. The continued expansion of major towns and cities in this region is conducive to rising levels of air pollution, mainly due to increasing numbers of public transport and private vehicles. Therefore, the focus on this thesis was to investigate the relationship between NO_2 concentrations, a measure of traffic-related air pollution, and ill health in West Central Scotland during the period 2006 to 2012. Health outcomes of interest included cardio-respiratory morbidity and mortality. Cardio-respiratory disease was chosen as the study outcome due to its well-established links with air pollution (Scoggins et al., 2004; Wang et al., 2009), since pollutants travel deep into the lungs when breathed, thus aggravating the respiratory tract.

Researchers are faced with many challenges when it comes to statistical modelling, especially when trying to estimate air pollutant concentrations at the appropriate spatial level to be used in conjunction available disease data (referred to as the change-of-support problem in the statistical literature). Therefore, the aims of this thesis were to firstly, develop an air pollutant model that produces accurate and spatially dense NO_2 concentrations, that also mitigates the statistical challenges faced, such as the change-of-support problem, when combining different sets of spatial data. Secondly, these predicted pollutant concentrations were used to investigate its association with cardio-respiratory ill health in West Central Scotland. Thirdly, estimates of the effect of NO_2 on ill health may be influenced by the type of NO_2 concentrations used, the choice of deprivation indicator, and the choice of spatial model, therefore, this thesis also explored the sensitivity of the NO_2-health relationship to these choices and suggested ways of accounting for uncertainty.

7.2 Estimating spatially representative NO_2 concentrations

The overarching aim of this thesis was to investigate the impact of NO_2 concentrations on cardio-respiratory disease, from 2006 to 2012 in West Central Scotland, thus requiring spatially representative air pollutant concentrations. The disease data were available at the data zone level, whereas NO_2 data were only available at the point and grid level. The misalignment in data between disease and air pollutant formed the basis for this chapter, which sought to estimate accurate, fine scale NO_2 concentrations at the desired spatial scale.

The NO_2 concentrations across West Central Scotland were available from two sources: measured concentrations from automatic monitors and diffusion tubes, and modelled concentrations from an atmospheric dispersion model. The measured concentrations were sparse at the data zone level in that there was not an air pollution monitor situated within every areal unit. This is why modelled concentrations are typically used in epidemiological studies, since they provide complete spatial coverage at a fine spatial scale. However, biases were inherent in these modelled concentrations, since they were estimated from a mathematical model that usually provides no measure of uncertainty. This chapter developed a geostatistical fusion model to combine both sets of data on NO_2 in order to predict NO_2 concentrations at a fine spatial scale that could then be used alongside the disease data.

The geostatistical fusion model developed in this thesis was an extension of the model developed by Berrocal et al. (2010b), and was novel because it utilised information from diffusion tubes, which have not been available for use in previous studies (Lawson et al., 2012), as well as utilising important covariate information. The automatic monitors only provide information at a limited number of locations, 16 in West Central Scotland, and utilising diffusion tubes increased this number by 230 locations. This in turn helped increase the efficacy of the geostatistical fusion model.

This chapter also studied the robustness of the fusion model in terms of comparing modelling within a Bayesian setting to a classical frequentist setting. Both models performed similarly when comparing bias and RMSPE, but the Bayesian model had wider and more appropriate prediction intervals, owing to the fact that modelling within a Bayesian framework allows for uncertainty in the estimated model parameters, whereas the frequentist approach assumes the estimated model parameters are fixed and known when making predictions. Furthermore, this chapter demonstrated an improvement in fine scale spatial prediction when the diffusion tube data were included with the commonly used automatic monitors. This improvement was due to the sheer increase in the number of spatial locations at which NO_2 concentrations were measured. This was shown in the comparison of a prediction model solely utilising the automatic monitors, and a prediction model solely using the diffusion tubes. The diffusion tube prediction model performed better than the automatic monitor prediction model, as it had lower bias (meaning the predictions were close to the true values), lower RMSPE, and predictions intervals that were not overly wide. However, there was no difference in prediction accuracy when both the automatic monitors and diffusion tube were used jointly to predict NO_2 concentrations. This is not surprising since it was the addition from the diffusion tubes that improved the prediction accuracy, not vice versa.

There are a number of ways in which the proposed geostatistical fusion model can be extended. One way is in terms of including a temporal resolution to the model. A

purely spatial study was conducted in this thesis due to the lack of variation between the NO_2 concentrations over the years, and also for the fact that the health study was at a yearly level and purely spatial. This is a limitation of the present study. Another limitation is in terms of temporal resolution, as the diffusion tube data are only available as monthly averages, whereas monitor data are available either hourly or daily. Therefore, predicting at finer temporal scales is not possible at present. In addition, it would be desirable to predict at a finer spatial resolution when used in conjunction with the disease data. Although disease data are available at the area level, they still require fine spatial resolutions since the size of a data zone depends on population density. Therefore, in the city where population density is greatest the data zones are extremely small and pollutant concentrations at 1km grid square are still too large to ensure each small data zone has a grid box centroid, which was used as the location point to assign each data zone a NO_2 value. Despite these limitations, the clear improvements in fine scale spatial prediction due to the novel methodology are incredibly important in taking the estimation of air pollutant concentrations forward for future research.

7.3 Application of estimated NO_2 concentrations to cardio-respiratory mortality data

The overarching aim of Chapter 5 was to apply the NO_2 concentrations developed from the previous chapter to cardio-respiratory mortality data in West Central Scotland to investigate whether long-term exposure to NO_2 concentrations increased the risk of cardio-respiratory mortality. While no substantial association was observed between exposure to NO_2 and cardio-respiratory mortality after adjustment for deprivation, this chapter provided a key insight into the importance of ensuring the pollutant-health effect is robust to the modelling choices made. Having numerous covariates or different statistical models (spatial in the present case) can lead to a variation estimated effect sizes. Variation in estimated effect sizes is often ignored in epidemiological studies, which tend to highlight their 'best' model depending on the model that minimised some goodness-of-fit criteria, such as the AIC or DIC. However, in air pollution and health studies, the estimated effect sizes are relatively small meaning, the overall conclusion of the research depends on the final model chosen.

This chapter investigated the sensitivity of the NO_2-health effect according to three factors: estimation of NO_2 concentrations, the choice of deprivation indicator, and the choice of spatial model. Then sought to provide an approach to utilising information from all models, rather than just presenting information from a single model. The proposed methodology utilised Bayesian model averaging (BMA) as a way of combining the results from all models into a single overall estimate that takes model uncertainty

into account. The main finding was that the NO_2-health effect was not robust to the choice of these three factors, as a wide range of effect sizes were observed. There was attenuation of the pollutant-heath effects when the NO_2 concentrations were estimated using the proposed geostatistical fusion model compared to when the modelled DEFRA concentrations were used. Furthermore, when BMA was performed on all models, the statistical model with the DEFRA concentrations contributed to the overall effect size more rather than the model with the proposed fusion model concentrations. The majority of spatial ecological studies in Scotland (and the UK) tend to use modelled concentrations due to their wide availability and ease of use, even though they are considered to contain biases. This suggests that, studies which use these types of modelled data are using data that are more aligned with residual disease. However, as discussed in Chapter 4, the DEFRA data were not as good at predicting measured pollutant concentrations compared to the proposed fusion model, which provides more evidence that the proposed estimated NO_2 concentrations are indeed a better representation of NO_2 levels in West Central Scotland compared to the estimated DEFRA concentrations.

As mentioned previously, data zones in West Central Scotland are relatively small in the city due to its high population density, meaning that not all data zones have an air pollutant value since they do not contain a grid box centroid. These data zones were assigned the closest grid box centroid to where the population density is greatest. However, this can still lead to exposure misclassification, as, technically, the data zone is an amalgamation of all grid boxes that cross the data zone boundary. More research is needed to ensure that areal units take a representative value that minimises exposure misclassification, which in turn would lead to less biased and less diluted effect sizes.

There was a lack of variation across years in the cardio-respiratory mortality data used in this thesis, meaning that the death counts had to be aggregated over a seven year period. This can be mitigated by upgrading to a larger spatial unit to try and increase the power of the study. However, there must be a trade-off between improving power by having a coarser spatial unit and a reduction in evidence of a causal relationship. The smaller the areal unit, the smaller the population within each area. This implies a more socially and demographically homogeneous area, which is more closely related to an individual level study.

7.4 Application of estimated NO_2 concentrations to cardio-respiratory incidence data

The focus of this chapter was to study the effect of NO_2 concentrations on the incidence of cardio-respiratory disease. This chapter builds on knowledge from the previous chapter by only utilising the spatial model that had the best performance, which was the spatial model using the Leroux specification to model residual spatial autocorrelation. This chapter is also similar to the previous chapter also in that it investigated whether the relationship between NO_2 concentrations and ill health was different in different age groups.

Since the outcome of this chapter was different to that in the previous chapter, analyses were conducted by comparing, in the first instance, both sets of estimated NO_2 concentrations and utilising all indicators of deprivation in case different indicators contributed to the overall NO_2-health effect size. In contrast to the previous chapter, the type of estimated NO_2 concentrations and the choice of deprivation indicator did not have a major impact on the estimated NO_2-health relationship. Again, no substantial association was observed; however, the overall effect size indicated a protective effect of NO_2, rather than a detrimental effect. The mortality and incidence data were relatively highly correlated, suggesting that the hospital admission data may have a different relationship with air pollution compared to the mortality data. It was not possible to assess this difference within this thesis due to time constraints, therefore, future work could aim to investigate this.

Age is an important consideration in air pollution and health studies, as exposure to air pollution may impact more vulnerable groups, such as children and the elderly, who may be more susceptible to its detrimental effects. This chapter found inconsistent results across the three age groups: younger ages (0-19 years), working ages (20-64 years), and elderly ($\geqslant$ 65 years). These different age groups encompassed a wide range of ages and thus, a wide range of morbidities. Previous studies found associations between air pollution and cardio-respiratory disease in young children and in the older population, but in this thesis no such relationships were found. It is possible that this was due to the age groups used here being too wide for any association to be noticeable. It was not possible in this thesis to construct narrower age groups, as smaller groups would lead to smaller numbers of cardio-respiratory disease within the groups and thus, less variation in the outcome. Therefore, there is scope for future studies in Scotland to add to the growing evidence that exposure to air pollution affects more vulnerable age groups by studying this further.

As with all air pollution and health studies, there is always a degree of exposure misclassification, and within this chapter, exposure misclassification was prominent.

The level of exposure misclassification may not be as high in cohort studies, since exposure is assigned individually and can, therefore, borrow information from other factors, such as information on where the individual spends the majority of their time. However, ecological studies infer associations at the group level, where every individual within an area is assigned the same level of exposure, which clearly is not plausible. It is important for future research to mitigate exposure misclassification when it comes to assigning exposure levels to younger and working age populations, since these groups of people are more likely to spend their time not in the home.

Deprivation has played a major role in the two pollutant-health studies conducted in this thesis. The majority of ecological studies control for deprivation using one or two variables that serve as a proxy for area level deprivation. In this thesis, each individual domain of the overall Scottish Index of Multiple Deprivation was included separately to investigate which deprivation indicators were the most important in the NO_2-health relationship. This can help inform future research about the deprivation indicators that should always be considered in an analysis. Bayesian model averaging was performed as a way of determining which indicators were the most important. There was consistency in the the deprivation indicator having the most influence between the two pollutant-health chapters; however, the indicators changed when stratified by age group. However, correlation between these indicators (income, employment, and education) was high, suggesting that these indicators can be used interchangeably as they essentially provide the same information. This is beneficial for future studies as it is not always possible to obtain data on the suggested important indicators; thus, knowing that there are a number of indicators that can be used eases the burden of lack of available data or information.

This thesis only made use of a single air pollutant, NO_2, since data were more widely available. This pollutant served as an indication for traffic-related pollution, and is widely studied due to its known health risks. However, the air people breathe is a complex mixture of a whole host of air pollutants. Ideally, multi-pollutant models should be considered, where care should be taken to ensure no two highly correlated pollutants are included in the same statistical model. Combining air pollutants into an overall air pollution index would be one way of ensuring air pollution is considered as a whole. However, it is equally important to investigate individual pollutants, since they are produced in different ways from different sources, and can help inform local government and policy on how to reduce pollutants by targeting specific sources, and thus improve overall air quality.

It is known that smoking is a major cause of both cardiovascular and respiratory disease, and is a cause of air pollution in the home. Smoking is also very strongly socially patterned, with higher levels of smoking in more deprived areas, and therefore

could be influential in all aspects of this thesis. However, as information on smoking is not routinely collected, it could not be accounted for in this thesis. It was also not possible to assess, separately, the potential impacts of socio-economic deprivation and smoking on the pollutant-health effects. There are many implications for not adjusting for smoking, such as biased and stronger pollutant-health effects, and perhaps also stronger deprivation-health effects which would be attenuated if smoking was included. However, smoking has been shown to be highly correlated with deprivation (Rushworth et al., 2014), which has been accounted for in these analyses. Future analyses could adopt the approaches suggested by Jackson et al. (2006) and Wakefield (2004), who suggest combining area-level disease and air pollution data with corresponding individual-level smoking data on smaller samples within a select number of areas. This could reduce the bias compared with a simple spatial ecological study, and also has the ability to increase the power of an individual-level only study. Although small areas, such as data zones and output areas, are typically constructed for use in a census, where data on smoking are still not routinely collected as part of this. Instead, data on smoking and other lifestyle factors are becoming more widely available through health surveys, meaning that these types of data could be incorporated into future analyses to utilise the aforementioned approaches.

Lastly, one of the main issues in this thesis and in other air pollution and health studies is that the uncertainty present in the air pollution concentrations is not fully taken into account. This thesis and the majority of studies treat the air pollutant concentrations, whether estimated from a geostatistical fusion model like here or utilise output from an atmospheric dispersion model, as the known and true value (Huang et al., 2015; Lee & Sarran, 2015; Rushworth et al., 2014). This could give rise to a biased estimate of the relationship between air pollution and health. This weakness can be overcome in a number of ways, one of which is to develop a joint Bayesian framework where the posterior distribution of the estimated air pollutant concentrations are directly fed forward into the health model, as demonstrated by Blangiardo et al. (2016). However, care has to be taken when utilising a joint modelling framework so that the health model does not infer the air pollution model, known as *feedback*, as it can artificially increase the precision of the estimates of the air pollutant effects (Blangiardo et al., 2016). This is therefore an avenue for future work for this thesis and future air pollutant and health studies.

7.5 Summary

This thesis involved three major pieces of research, which enhance and contribute to the existing literature surrounding air pollution and health. It provides a new approach to estimating pollutant concentrations at a fine spatial scale that can then be used to investigate its effects on human ill health, while determining which factors in an analysis

have the most influence on the resulting estimated pollutant-health association. The methodology developed in this thesis can be adopted in future studies, and should lead to further research into this important public health topic, and provides a more accurate approach to estimating air pollutant concentrations, with the ability to combine results from multiple models. The novel air pollutant model, given by a geostatistical fusion model, allows concentrations measured at different spatial resolutions, such as the point and grid level, to be combined to produce a more accurate representation of air pollutant concentrations at the desired spatial resolution (grid level in the present case) that are more spatially dense compared to using point-level concentrations from air pollution monitors, or solely using initial grid-level concentrations that are biased and have no measure of uncertainty. Furthermore, this geostatistical fusion model can be extended to include a temporal domain when there is variation in concentrations across the study period.

This thesis highlighted that the estimated pollutant-health effect depends on how air pollutant concentrations are measured, the measure of deprivation used, and finally, the spatial model used. It demonstrated how selecting only one model ignored any uncertainty present.

While no substantial association was found between NO_2 concentrations and cardio-respiratory disease, results from this thesis add to the body of evidence surrounding air pollution and ill health. With regards to policy and legislation, it is important to continually improve public health through closer integration of air quality and climate change policies, even when no consistent associations are observed. Tackling air pollution as part of the UK and Scottish Climate Change Acts could result in saving $24 billion by 2050. However, short-term solutions are needed to meet current EU air quality thresholds in Scotland.

There are a number of strategies that can be implemented to tackle the damaging effects of air pollution on ill health. These mainly focus on emission control and include strategies, such as low emission zones, transport planning at the local level (e.g., bus management arrangements), and adopting low carbon or hydrogen-electric vehicles. However, while most studies focus on outdoor air pollution, the highest level of exposure occurs within the household, implying a greater focus on energy-efficient housing would also be beneficial. Efforts to improve air quality and the built environment have the potential to improve the health of the population as a whole.

'It seems to me that the natural world is the greatest source of excitement; the greatest source of visual beauty; the greatest source of intellectual interest. It is the greatest source of so much in life that makes life worth living.'

— Sir David Attenborough

Appendix A

NO_2 predicted pollution maps

This appendix details predicted NO_2 pollution maps across West Central Scotland from the Bayesian Geostatistical fusion model proposed in Section 6.3.

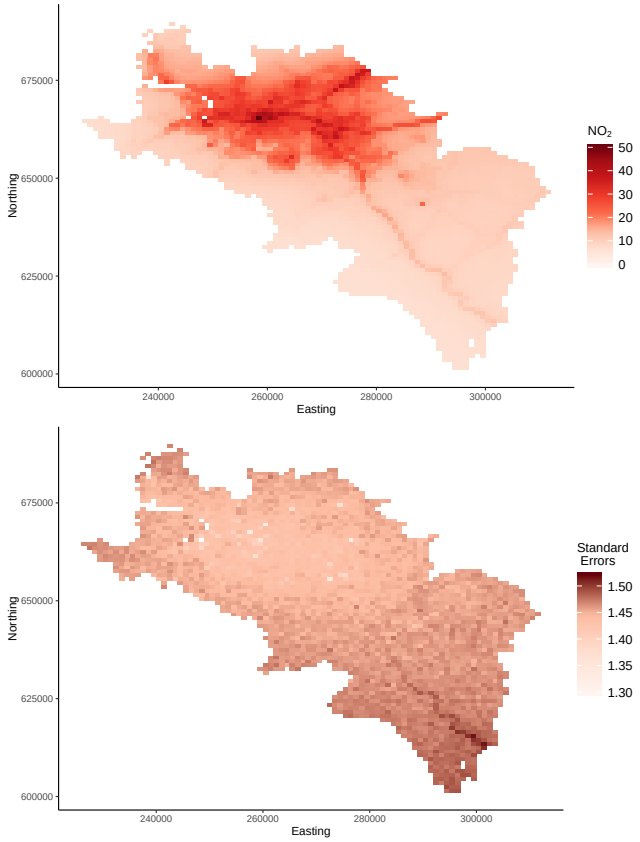

Figure A.1: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2007.*

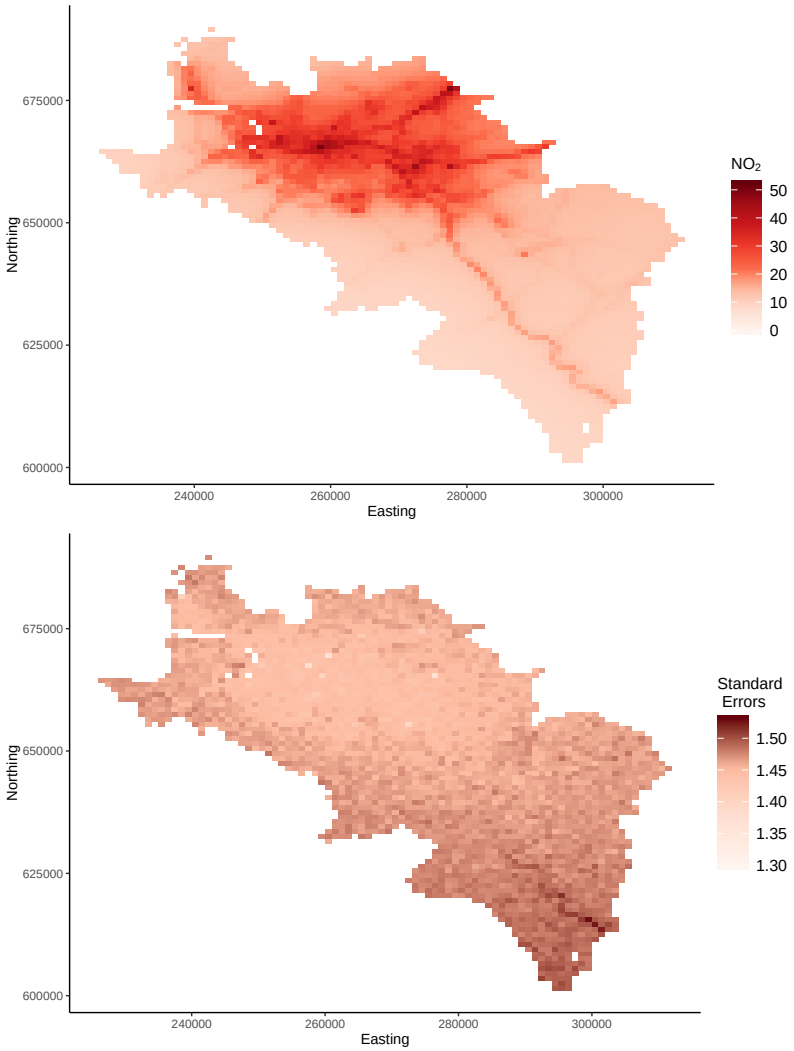

Figure A.2: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2008.*

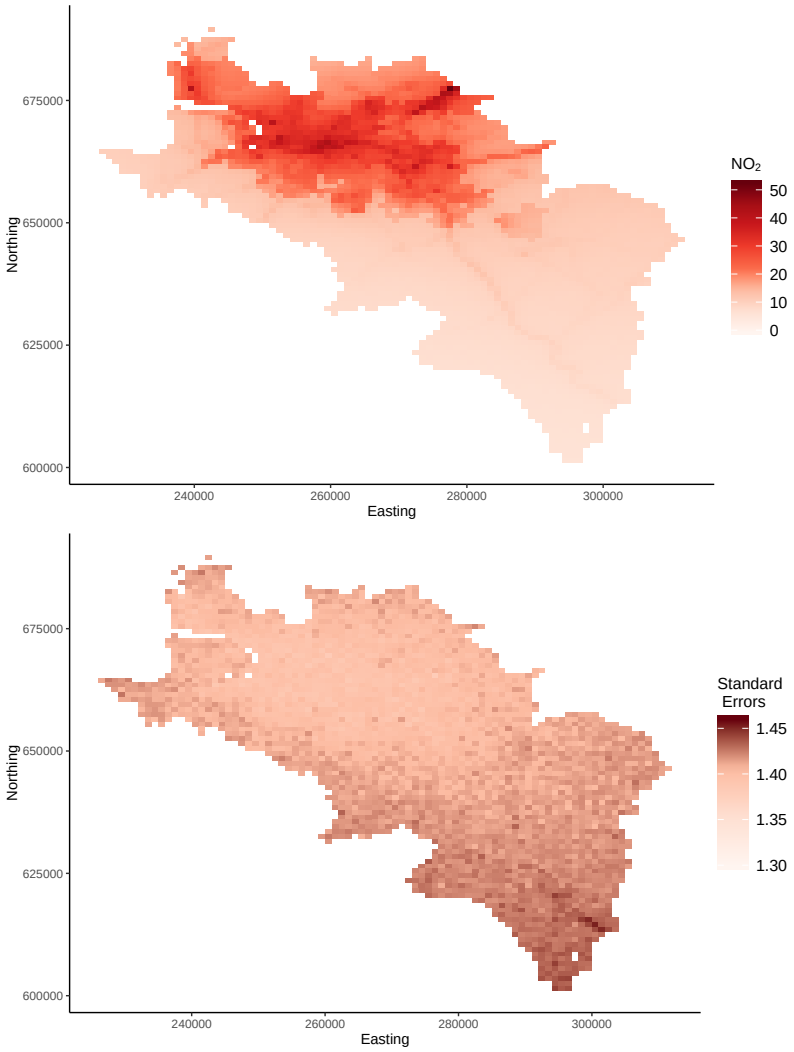

Figure A.3: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2009.*

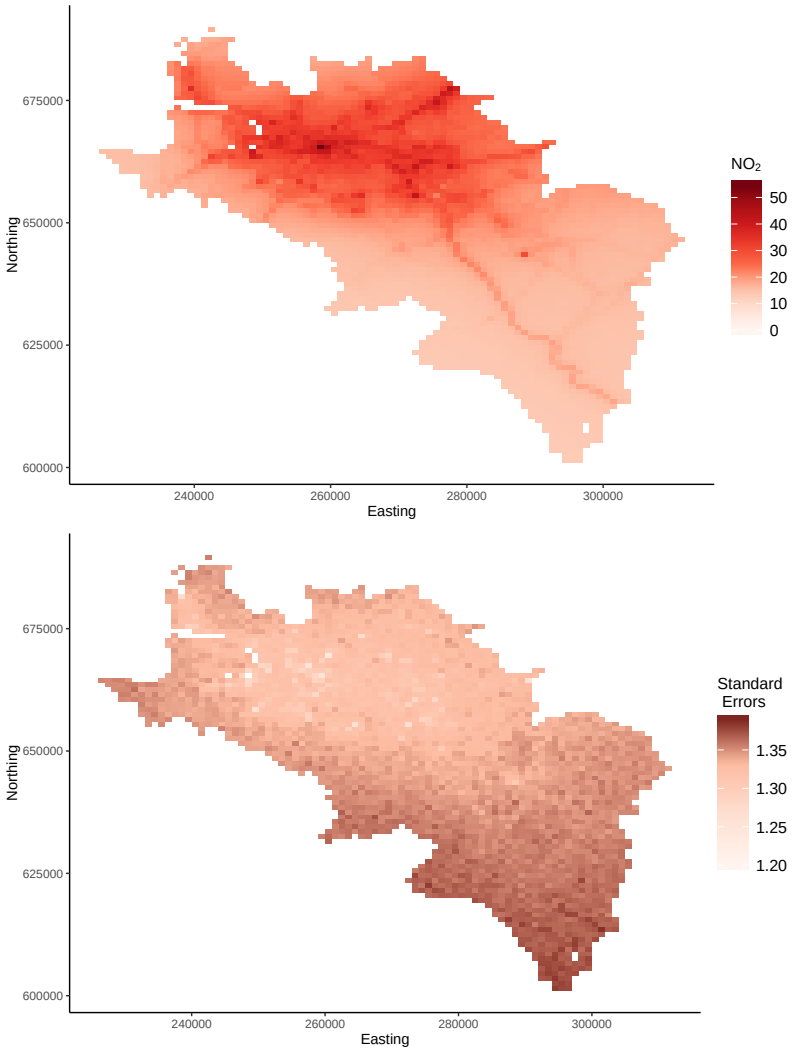

Figure A.4: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2010.*

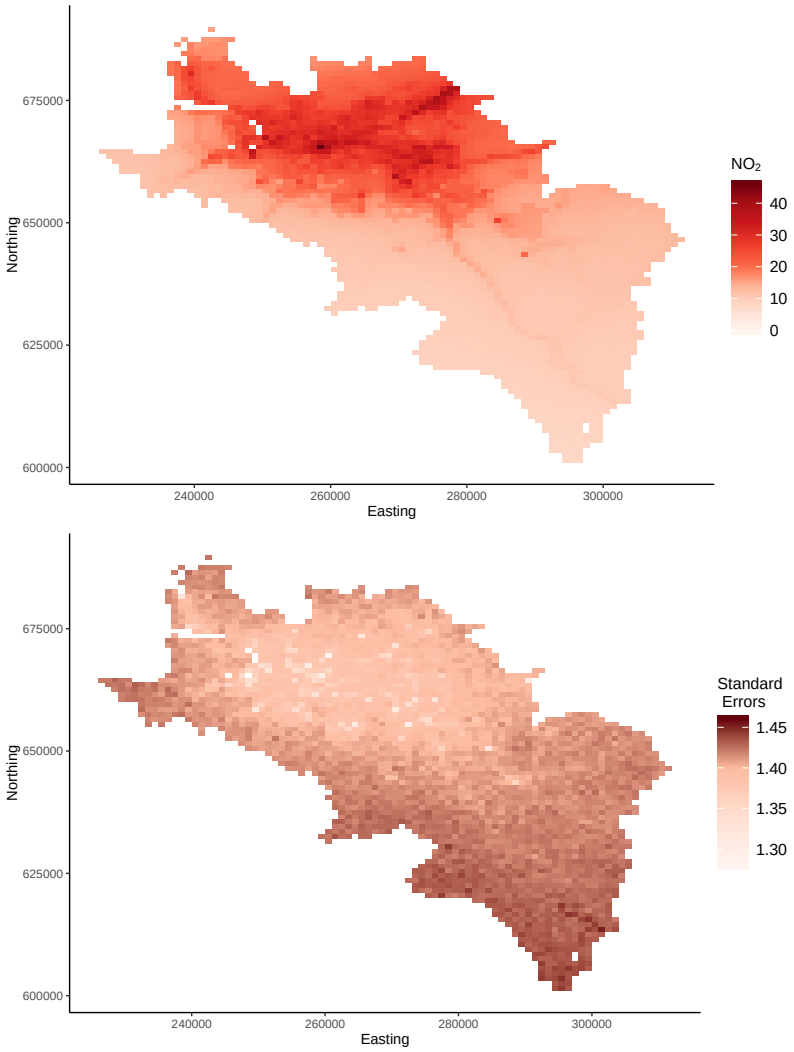

Figure A.5: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2011.*

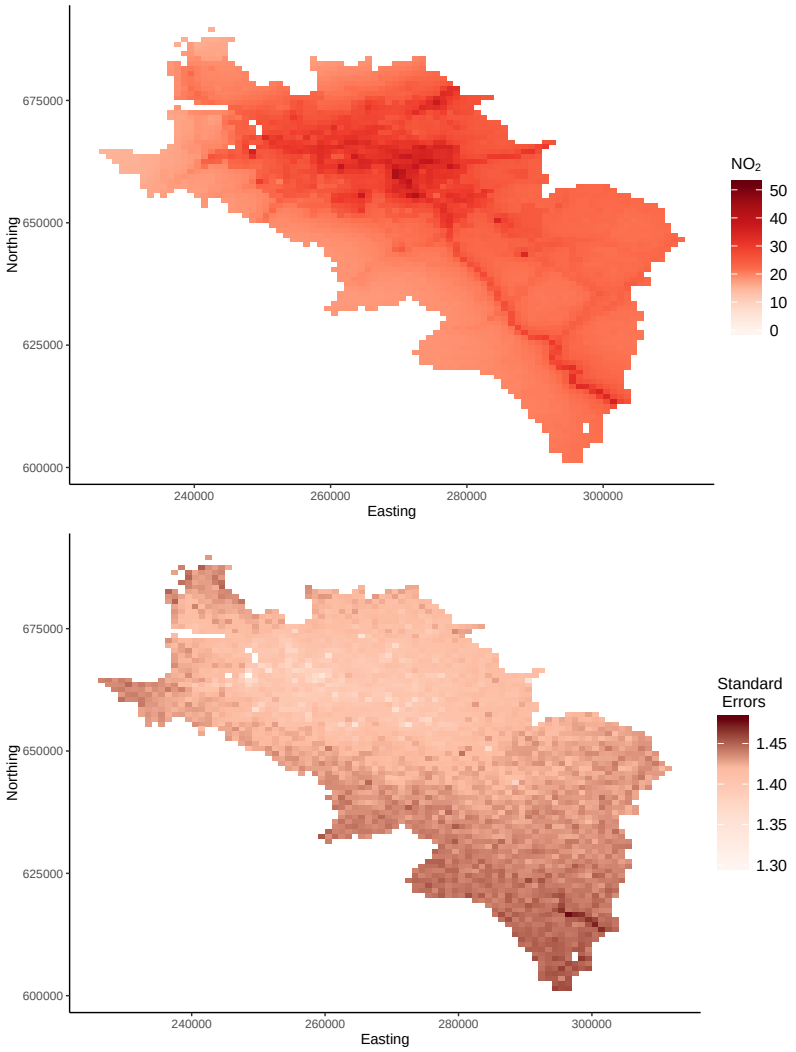

Figure A.6: *Spatial map of predicted NO_2 concentrations and standard errors from Model 9 across West Central Scotland in 2012.*

References

Agius, R. M., Cohen, G. R., Beverland, I., Elton, R. A., Lee, R., & Boyd, J. (2002). Epidemiological Study of Susceptibility to Cardiorespiratory Death from Particulate Air Pollution. *The Annals of Occupational Hygiene*, *46*, 452–455. doi:10.1093/annhyg/mef717.

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*.

Anderson, H. (2009). Air pollution and mortality: A history. *Atmospheric Environment*, *43*, 142–152. doi:10.1016/j.atmosenv.2008.09.026.

Anderson, H. R., Atkinson, R. W., Bremner, S. A., & Marston, L. (2003). Particulate air pollution and hospital admissions for cardiorespiratory diseases: are the elderly at greater risk? *The European Respiratory Journal. Supplement*, *40*, 39s–46s.

Armstrong, B. G. (1998). Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and Environmental Medicine*, *55*, 651–656. doi:10.1136/oem.55.10.651.

Atkinson, R. W., Anderson, H. R., Sunyer, J., Ayres, J., Baccini, M., Vonk, J. M., Boumghar, A., Forastiere, F., Forsberg, B., Touloumi, G., Schwartz, J., & Katsouyanni, K. (2001). Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project. Air Pollution and Health: a European Approach. *American Journal of Respiratory and Critical Care Medicine*, *164*, 1860–6. doi:10.1164/ajrccm.164.10.2010138.

Atkinson, R. W., Carey, I. M., Kent, A. J., van Staa, T. P., Anderson, H. R., & Cook, D. G. (2013). Long-term exposure to outdoor air pollution and incidence of cardiovascular diseases. *Epidemiology (Cambridge, Mass.)*, *24*, 44–53. doi:10.1097/EDE.0b013e318276ccb8.

Bailey, L., Vardulaki, K., Langham, J., & Chandramohan, D. (2005). *Introduction to Epidemiology*. Understanding Public Health. Maidenhead: Open University Press.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Barceló, M. A., Saez, M., & Saurina, C. (2009). Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. *The Science of the total environment*, *407*, 5501–23. doi:10.1016/j.scitotenv.2009.07.028.

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

Beatty, T. K., & Shimshack, J. P. (2014). Air pollution and children's respiratory health: A cohort analysis. *Journal of Environmental Economics and Management*, *67*, 39–57. doi:10.1016/j.jeem.2013.10.002.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A. S., Lanki, T., Turunen, A., Oftedal, B., Nystad, W., Nafstad, P., De Faire, U., Pedersen, N. L., Östenson, C.-G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K. T., Overvad, K., Ellermann, T., Eeftens, M., Peeters, P. H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Hampel, R., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Künzli, N., Schindler, C., Schikowski, T., Adam, M., Phuleria, H., Vilier, A., Clavel-Chapelon, F., Declercq, C., Grioni, S., Krogh, V., Tsai, M.-Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Brunekreef, B., & Hoek, G. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet*, *383*, 785–95. doi:10.1016/S0140-6736(13)62158-3.

Beelen, R., Voogt, M., Duyzer, J., Zandveld, P., & Hoek, G. (2010). Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmospheric Environment*, *44*, 4614–4621. doi:10.1016/j.atmosenv.2010.08.005.

Bennett, O., Kandala, N.-B., Ji, C., Linnane, J., & Clarke, A. (2014). Spatial variation of heart failure and air pollution in Warwickshire, UK: an investigation of small scale variation at the ward-level. *BMJ open*, *4*, e006028. doi:10.1136/bmjopen-2014-006028.

Bernstein, J. a., Alexis, N., Barnes, C., Bernstein, I. L., Nel, A., Peden, D., Diaz-Sanchez, D., Tarlo, S. M., & Williams, P. B. (2004). Health effects of air pollution. *The Journal of Allergy and Clinical Immunology*, *114*, 1116–23. doi:10.1016/j.jaci.2004.08.030.

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010a). A bivariate space-time downscaler under space and time misalignment. *Annals of Applied Statistics*, *4*, 1942–1975. doi:10.1214/10-AOAS351.

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010b). A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of agricultural, biological, and environmental statistics*, *15*, 176–197. doi:10.1007/s13253-009-0004-z.

Bertazzon, S., Johnson, M., Eccles, K., & Kaplan, G. G. (2015). Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and Spatio-temporal Epidemiology*, *14-15*, 9–21. doi:10.1016/j.sste.2015.06.002.

Besag, J., York, J., & Mollie, a. (1991). Bayesian image-restoration, with 2 applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*, 1–20. doi:10.1007/BF00116466.

Beverland, I., Carder, M., Cohen, G., Heal, M., & Agius, R. (2014a). Associations between short/medium-term variations in black smoke air pollution and mortality in the Glasgow conurbation, UK. *Environment International*, *62*, 126–132. doi:10.1016/j.envint.2013.01.002.

Beverland, I., Carder, M., Cohen, G., Heal, M., & Agius, R. (2014b). Associations between short/medium-term variations in black smoke air pollution and mortality in the Glasgow conurbation, UK. *Environment International*, *62*, 126–132. doi:10.1016/j.envint.2013.01.002.

Beverland, I. J., Cohen, G. R., Heal, M. R., Carder, M., Yap, C., Robertson, C., Hart, C. L., & Agius, R. M. (2012a). A comparison of short-term and long-term air pollution exposure associations with mortality in two cohorts in Scotland. *Environmental Health Perspectives*, *120*, 1280–5. doi:10.1289/ehp.1104509.

Beverland, I. J., Robertson, C., Yap, C., Heal, M. R., Cohen, G. R., Henderson, D. E. J., Hart, C. L., & Agius, R. M. (2012b). Comparison of models for estimation of long-term exposure to air pollution in cohort studies. *Atmospheric Environment*, *62*, 530–539. doi:10.1016/j.atmosenv.2012.08.001.

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2013). Applied spatial data analysis with R. In *Use R!*. Springer-Verlag New York. doi:10.1007/978-1-4614-7618-4.

Blangiardo, M., Finazzi, F., & Cameletti, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, *0*, 1–12. doi:10.1016/j.sste.2016.03.001.

Brookes, D., Steadman, J., Grice, S., Kent, A., Walker, H., Cooke, S., Vincent, K., Lingard, J., Bush, T., & Abbott, J. (2011). *UK modelling under*

the Air Quality Directive (2008/50/EC) for 2010 covering the following air quality pollutants: SO_2, NO_x, NO_2, PM_{10}, $PM_{2.5}$, lead, benzene, CO and ozone. AEAT/ENV/R/3215 Issue 1.. Technical Report Department for Environment, Food & Rural Affairs, UK (No. 3215). URL: http://uk-air.defra.gov.uk/reports/cat09/1204301513{_}AQD2010mapsrep{_}master{_}v0.pdf.

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brunekreef, B. (2007). Health effects of air pollution observed in cohort studies in Europe. *Journal of Exposure Science and Environmental Epidemiology*, *17*, S61–5. doi:10.1038/sj.jes.7500628.

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *Lancet*, *360*, 1233–42. doi:10.1016/S0140-6736(02)11274-8.

Bruno, F., Cocchi, D., Greco, F., & Scardovi, E. (2013). Spatial reconstruction of rainfall fields from rain gauge and radar data. *Stochastic Environmental Research and Risk Assessment*, *28*, 1235–1245. doi:10.1007/s00477-013-0812-0.

Carder, M., McNamee, R., Beverland, I., Elton, R., Cohen, G. R., Boyd, J., Van Tongeren, M., & Agius, R. M. (2010). Does deprivation index modify the acute effect of black smoke on cardiorespiratory mortality? *Occupational and Environmental Medicine*, *67*, 104–10. doi:10.1136/oem.2008.044602.

Carder, M., McNamee, R., Beverland, I., Elton, R., Van Tongeren, M., Cohen, G. R., Boyd, J., Macnee, W., & Agius, R. M. (2008). Interacting effects of particulate pollution and cold temperature on cardiorespiratory mortality in Scotland. *Occupational and Environmental Medicine*, *65*, 197–204. doi:10.1136/oem.2007.032896.

Carstairs, V. (2001). Socio-economic factors at areal level and their relationship with health. In *Spatial Epidemiology* (pp. 51–67). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198515326.003.0004.

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., & Forastiere, F. (2013). Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environmental health perspectives*, *121*, 324–31. doi:10.1289/ehp.1205862.

Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., de Faire, U., Erbel, R., Eriksen, K. T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jöckel, K.-H., Korek, M., Lanki, T., Leander, K., Magnusson, P. K. E., Migliore, E., Ostenson, C.-G., Overvad, K., Pedersen, N. L., J, J. P., Penell, J., Pershagen,

G., Pyko, A., Raaschou-Nielsen, O., Ranzi, A., Ricceri, F., Sacerdote, C., Salomaa, V., Swart, W., Turunen, A. W., Vineis, P., Weinmayr, G., Wolf, K., de Hoogh, K., Hoek, G., Brunekreef, B., & Peters, A. (2014). Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ (Clinical research ed.)*, *348*. doi:10.1136/bmj.f7412.

Ciocco, A., & Thompson, D. J. (1961). A follow-up of Donora ten years after: methodology and findings. *American journal of public health*, *51*, 155–64.

Clayton, D. G., Bernardinelli, L., & Montomoli, C. (1993). Spatial correlation in ecological analysis. *International journal of epidemiology*, *22*, 1193–202.

Committee on Environmenta Health (2004). Ambient Air Pollution: Health Hazards to Children. *Pediatrics*, *114*, 1699 LP – 1707.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: a gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, *91*, 121–36. doi:10.1080/00223890802634175.

Cressie, N. (1993). *Statistics for spatial data, revised ed* volume 900. Wiley New York.

Crouse, D. L., Ross, N. A., & Goldberg, M. S. (2009). Double burden of deprivation and high concentrations of ambient air pollution at the neighbourhood scale in Montreal, Canada. *Social science & Medicine*, *69*, 971–81. doi:10.1016/j.socscimed.2009.07.010.

Dab, W., Medina, S., Quénel, P., Le Moullec, Y., Le Tertre, a., Thelot, B., Monteil, C., Lameloise, P., Pirard, P., Momas, I., Ferry, R., & Festy, B. (1996). Short term respiratory health effects of ambient air pollution: results of the APHEA project in Paris. *Journal of epidemiology and community health*, *50 Suppl 1*, s42–6.

Dalgaard, P. (2008). *Introductory Statistics with R*. Springer Science & Business Media.

Department for the Environment, Food and Rural Affairs (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland, publisher = The Stationery Office.

Department for the Environment, Food and Rural Affairs (2015). Air Quality Plans for the achievement of EU air quality limit values for nitrogen dioxide (NO2) in the UK. URL: http://uk-air.defra.gov.uk/assets/documents/no2ten/110921_UK_overview_document.pdf.

Dibben, C., & Clemens, T. (2015). Place of work and residential exposure to ambient air pollution and birth outcomes in Scotland, using geographically fine pollution climate mapping estimates. *Environmental Research*, *140*, 535–541. doi:10.1016/j.envres.2015.05.010.

Diggle, P., & Ribeiro, P. J. (2007). *Model-based geostatistics*. New York: Springer. doi:10.1007/978-0-387-48536-2.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., & Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine*, *329*, 1753–1759.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*, 89–121.

Elliott, P., & Savitz, D. A. (2008). Design issues in small-area studies of environment and health. *Environmental health perspectives*, *116*, 1098–104. doi:10.1289/ehp.10817.

Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J., & Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, *73*, 702–10.

Elliott, P., Shaddick, G., Wakefield, J. C., de Hoogh, C., & Briggs, D. J. (2007). Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax*, *62*, 1088–1094. doi:10.1136/thx.2006.076851.

Faraway, J. J. (2004). *Linear Models with R*. CRC Press.

Fernández-Somoano, A., Hoek, G., & Tardon, A. (2013). Relationship between area-level socioeconomic characteristics and outdoor NO2 concentrations in rural and urban areas of northern Spain. *BMC public health*, *13*, 71. doi:10.1186/1471-2458-13-71.

Firket, J. (1936). Fog along the Meuse Valley, . *32*, 1192–1197.

Fischer, P., Hoek, G., Brunekreef, B., Verhoeff, a., & van Wijnen, J. (2003). Air pollution and mortality in the Netherlands: are the elderly more at risk? *European Respiratory Journal*, *21*, 34S–38s. doi:10.1183/09031936.03.00402503.

Forastiere, F., Stafoggia, M., Tasco, C., Picciotto, S., Agabiti, N., Cesaroni, G., & Perucci, C. A. (2007). Socioeconomic status, particulate air pollution, and daily mortality: differential exposure or differential susceptibility. *American Journal of Industrial Medicine*, *50*, 208–216. doi:10.1002/ajim.20368.

Fuentes, M., & Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, *61*, 36–45. doi:10.1111/j.0006-341X.2005.030821.x.

Fuentes, M., Reich, B., & Lee, G. (2008). Spatial–temporal mesoscale modeling of rainfall intensity using gage and radar data. *The Annals of Applied Statistics*, *2*, 1148–1169.

Gelfand, A., & Sahu, S. (2010). Combining monitoring data and computer model output in assessing environmental exposure. In A. O'Hagan, & M. West (Eds.), *Handbook of Applied Bayesian Analysis* (pp. 458—-510). Oxford, UK: Oxford University Press.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., & Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, *13*, 263–312. doi:10.1007/BF02595775.

Gelfand, a. E., Zhu, L., & Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics (Oxford, England)*, *2*, 31–45. doi:10.1093/biostatistics/2.1.31.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierachical Models. *Bayesian Analysis*, *1*, 515 – 533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis*. (2nd ed.). Chapman & Hall/CRC.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 1–20. doi:10.1007/s11222-013-9416-2.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*, 457–472.

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

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian Statstics 4*. Clarendon Press, Oxford, UK.

Gittins, M., McNamee, R., Carder, M., Beverland, I., & Agius, R. M. (2013). Has the short-term effect of black smoke exposure on pneumonia mortality been underestimated because hospitalisation is ignored: findings from a case-crossover study. *Environmental Heath*, *12*, 97. doi:10.1186/1476-069X-12-97.

Goodman, A., Wilkinson, P., Stafford, M., & Tonne, C. (2011). Characterising socio-economic inequalities in exposure to air pollution: a comparison of socio-economic markers and scales of measurement. *Health & place*, *17*, 767–74. doi:10.1016/j.healthplace.2011.02.002.

Gotway, C. A., & Young, L. J. (2002). Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, *97*, 632–648. doi:10.1198/016214502760047140.

Greenland, S., & Morgenstern, H. (1989). Ecological bias, confounding, and effect modification. *International journal of epidemiology*, *18*, 269–74.

Haining, R., Li, G., Maheswaran, R., Blangiardo, M., Law, J., Best, N., & Richardson, S. (2010). Inference from ecological models: estimating the relative risk of stroke from air pollution exposure using small area data. *Spatial and spatio-temporal epidemiology*, *1*, 123–31. doi:10.1016/j.sste.2010.03.006.

Haneuse, S. J.-P. A., & Wakefield, J. C. (2007). Hierarchical models for combining ecological and case-control data. *Biometrics*, *63*, 128–36. doi:10.1111/j.1541-0420.2006.00673.x.

Haneuse, S. J.-P. A., & Wakefield, J. C. (2008). The Combination of Ecological and Case-Control Data. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *70*, 73–93.

Hansell, A., Ghosh, R. E., Blangiardo, M., Perkins, C., Vienneau, D., Goffe, K., Briggs, D., & Gulliver, J. (2016). Historic air pollution exposure and long-term mortality risks in England and Wales: prospective longitudinal cohort study. *Thorax*, *71*, 330–338. doi:10.1136/thoraxjnl-2015-207111.

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

Havard, S., Deguen, S., Zmirou-Navier, D., Schillinger, C., & Bard, D. (2009). Traffic-related air pollution and socioeconomic status: a spatial autocorrelation study to assess environmental equity on a small-area scale. *Epidemiology*, *20*, 223–30. doi:10.1097/EDE.0b013e31819464e1.

Heidelberger, P., & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, *24*, 233–245. doi:10.1145/358598.358630.

Heidelberger, P., & Welch, P. D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research*, *31*, 1109–1144.

Hodges, J. S., & Reich, B. J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician*, *64*, 325–334. doi:10.1198/tast.2010.10052.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*, 7561–7578. doi:10.1016/j.atmosenv.2008.05.057.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., & van den Brandt, P. A. (2002). Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *The Lancet*, *360*, 1203–1209.

Hoek, G., Brunekreef, B., Verhoeff, A., Wijnen, J. V., & Fischer, P. (2000). Daily Mortality and Air Pollution in the Netherlands. *Journal of the Air & Waste Management Association*, *50*, 1380–1389. doi:10.1080/10473289.2000.10464182.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*, 382–417. doi:10.2307/2676803.

Hu, Z., Liebens, J., & Rao, K. R. (2008). Linking stroke mortality with air pollution, income, and greenness in northwest Florida: an ecological geographical study. *International Journal of Health Geographics*, *7*. doi:10.1186/1476-072X-7-20.

Huang, F., Chen, R., Shen, Y., Kan, H., & Kuang, X. (2016). The Impact of the 2013 Eastern China Smog on Outpatient Visits for Coronary Heart Disease in Shanghai, China. *International Journal of Environmental Research and Public Health*, *13*, 627. doi:10.3390/ijerph13070627.

Huang, G., Lee, D., & Scott, E. M. (2015). An integrated bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland. *Spatial and Spatio-temporal Epidemiology*, . doi:http://dx.doi.org/10.1016/j.sste.2015.09.002.

Huang, Y., Dominici, F., & Bell, M. L. (2005). Bayesian hierarchical distributed lag models for summer ozone exposure and cardio-respiratory mortality. *Environmetrics*, *16*, 547–562. doi:10.1002/env.721.

Hughes, J., & Cui, X. (1-16-2015). *ngspatial: Fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data*. Minneapolis, MN. R package version 1.0-5.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society Series B*, *75*, 139–159. URL: http://EconPapers.repec.org/RePEc:bla:jorssb:v:75:y:2013:i:1:p:139-159.

Jackson, C., Best, Nicky, & Richardson, S. (2008). Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*, 159–178. doi:10.1111/j.1467-985X.2007.00500.x.

Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in medicine*, *25*, 2136–59. doi:10.1002/sim.2370.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* volume 103 of *Springer Texts in Statistics*. New York, NY: Springer New York. doi:10.1007/978-1-4614-7138-7.

Janes, H., Dominici, F., & Zeger, S. L. (2007). Trends in Air Pollution and Mortality: an approach to the assessment of unmeasured confounding. *Epidemiology*, *18*, 416–423. doi:10.1097/EDE.0b013e31806462e9.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. (2005a). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, *15*, 185–204. doi:10.1038/sj.jea.7500388.

Jerrett, M., Burnett, R. T., Brook, J., Kanaroglou, P., Giovis, C., Finkelstein, N., & Hutchison, B. (2004). Do socioeconomic characteristics modify the short term association between air pollution and mortality? Evidence from a zonal time series in Hamilton, Canada. *Journal of Epidemiology and Community Health*, *58*, 31–40. doi:10.1136/jech.58.1.31.

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., & Thun, M. J. (2005b). Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology*, *16*, 727–736. doi:10.1097/01.ede.0000181630.15826.7d.

Jerrett, M., Buzzelli, M., Burnett, R. T., & DeLuca, P. F. (2005c). Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Social Science & Medicine*, *60*, 2845–63. doi:10.1016/j.socscimed.2004.11.006.

Jerrett, M., Finkelstein, M. M., Brook, J. R., Arain, M. A., Kanaroglou, P., Stieb, D. M., Gilbert, N. L., Verma, D., Finkelstein, N., Chapman, K. R., & Sears, M. R. (2009). A Cohort Study of Traffic-related Air Pollution and Mortality in Toronto, Canada. *Environmental Health Perspectives*, *772*, 772–777. doi:10.1289/ehp.11533.

Johnson, S., Bobb, J. F., Ito, K., Savitz, D. A., Elston, B., Shmool, J. L., Dominici, F., Ross, Z., Clougherty, J. E., & Matte, T. (2016). Ambient Fine Particulate Matter, Nitrogen Dioxide, and Preterm Birth in New York City. *Environmental Health Perspectives*, *124*, 1283–1290. doi:10.1289/ehp.1510266.

Katsouyanni, K., & Schwartz, J. (1996). Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol., . (pp. 1030–1038). doi:10.1183/09031936.95.08061030.

Kinney, P. L., & Ozkaynak, H. (1991). Associations of daily mortality and air pollution in Los Angeles County. *Environmental Research*, *54*, 99–120.

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*, 119–139. doi:doi:10.2307/3006914.

Laden, F., Neas, L. M., Dockery, D. W., & Schwartz, J. (2000). Association of fine particulate matter from different sources with daily mortality in six U.S. cities [In Process Citation]. *Environmental Health Perspectives*, *108*, 941–947. doi:10.1289/ehp.00108941.

Larrieu, S., Jusot, J.-F., Blanchard, M., Prouvost, H., Declercq, C., Fabre, P., Pascal, L., Tertre, A. L., Wagner, V., & Rivière, S. (2007). Short term effects of air pollution on hospitalizations for cardiovascular diseases in eight French cities: The PSAS program. *Science of The Total Environment*, *387*, 105–112. doi:10.1016/j.scitotenv.2007.07.025.

Laurent, O., Bard, D., Filleul, L., & Segala, C. (2007). Effect of socioeconomic status on the relationship between atmospheric pollution and mortality. *Journal of Epidemiology and Community Health*, *61*, 665–75. doi:10.1136/jech.2006.053611.

Laurent, O., Pedrono, G., Segala, C., Filleul, L., Havard, S., Deguen, S., Schillinger, C., Rivière, E., & Bard, D. (2008). Air pollution, asthma attacks, and socioeconomic deprivation: a small-area case-crossover study. *American Journal of Epidemiology*, *168*, 58–65. doi:10.1093/aje/kwn087.

Lawson, A. B., Choi, J., Cai, B., Hossain, M., Kirby, R. S., & Liu, J. (2012). Bayesian 2-Stage Space-Time Mixture Modeling With Spatial Misalignment of the Exposure in Small Area Health Data. *Journal of Agricultural, Biological, and Environmental Statistics*, *17*, 417–441. doi:10.1007/s13253-012-0100-3.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology*, *2*, 79–89. doi:10.1016/j.sste.2011.03.001.

Lee, D. (2012). Using spline models to estimate the varying health risks from air pollution across Scotland. *Statistics in Medicine*, *31*, 3366–78. doi:10.1002/sim.5420.

Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, *55*, 1–24. URL: http://www.jstatsoft.org/v55/i13/.

Lee, D., Ferguson, C., & Mitchell, R. (2009). Air pollution and health in Scotland: a multicity study. *Biostatistics*, *10*, 409–23. doi:10.1093/biostatistics/kxp010.

Lee, D., & Mitchell, R. (2012). Boundary detection in disease mapping studies. *Biostatistics (Oxford, England)*, *13*, 415–26. doi:10.1093/biostatistics/kxr036.

Lee, D., & Mitchell, R. (2014). Controlling for localised spatio-temporal autocorrelation in long-term air pollution and health studies. *Statistical methods in medical research*, (pp. 1–19). doi:10.1177/0962280214527384.

Lee, D., Rushworth, A., & Sahu, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, *70*, 419–429. doi:10.1111/biom.12156.

Lee, D., & Sarran, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics*, . doi:10.1002/env.2348.

Lee, D., & Shaddick, G. (2010). Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics*, *66*, 1238–46. doi:10.1111/j.1541-0420.2009.01376.x.

Leroux, B., Lei, X., & Breslow, N. (1999). Statistical Models in Epidemiology, the Environment and Clinical Trials. In M. Halloran, & D. Berry (Eds.), *Estimation of disease rates in small areas: A new mixed model for spatial dependence* (pp. 135–178). New York: Springer-Verlag. doi:10.1007/978-1-4612-1284-3.

Leyland, A. H., Dundas, R., McLoone, P., & Boddy, F. A. (2007a). Cause-specific inequalities in mortality in Scotland: two decades of change. A population-based study. *BMC public health*, *7*, 172. doi:10.1186/1471-2458-7-172.

Leyland, A. H., Dundas, R., McLoone, P., & Boddy, F. A. (2007b). Inequalities in mortality in {Scotland}, 1981-2001. Occasional Paper series no. 16. MRC Social and Public Health Science Unit.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution*, *3*, 112–115. doi:10.1111/j.2041-210X.2011.00131.x.

Liverani, S., Lavigne, A., & Blangiardo, M. (2016). Modelling collinear and spatially correlated data. *Spatial and Spatio-temporal Epidemiology*, *18*, 63–73. doi:10.1016/j.sste.2016.04.003.

Mackenbach, J. P., Kunst, A. E., Cavelaars, A. E., Groenhof, F., & Geurts, J. J. (1997). Socioeconomic inequalities in morbidity and mortality in western Europe. The EU Working Group on Socioeconomic Inequalities in Health. *Lancet*, *349*, 1655–9.

Mackenbach, J. P., Stirbu, I., Roskam, A.-J. R., Schaap, M. M., Menvielle, G., Leinsalu, M., & Kunst, A. E. (2008). Socioeconomic Inequalities in Health in 22 European Countries. *New England Journal of Medicine*, *358*, 2468–2481. doi:10.1056/NEJMsa0707519.

MacNab, Y. C., Kmetic, A., Gustafson, P., & Sheps, S. (2006). An innovative application of Bayesian disease mapping methods to patient safety research: a Canadian adverse medical event study. *Statistics in Medicine*, *25*, 3960–3980. doi:10.1002/sim.2507.

Maheswaran, R., Haining, R. P., Brindley, P., Law, J., Pearson, T., Fryers, P. R., Wise, S., & Campbell, M. J. (2005a). Outdoor air pollution and stroke in Sheffield, United Kingdom: a small-area level geographical study. *Stroke*, *36*, 239–43. doi:10.1161/01.STR.0000151363.71221.12.

Maheswaran, R., Haining, R. P., Brindley, P., Law, J., Pearson, T., Fryers, P. R., Wise, S., & Campbell, M. J. (2005b). Outdoor air pollution, mortality, and hospital admissions from coronary heart disease in Sheffield, UK: a small-area level ecological study. *European Heart Journal*, *26*, 2543–2549. doi:10.1093/eurheartj/ehi457.

Maheswaran, R., Haining, R. P., Pearson, T., Law, J., Brindley, P., & Best, N. G. (2006). Outdoor NOx and stroke mortality: adjusting for small area level smoking prevalence using a Bayesian approach. *Statistical methods in medical research*, *15*, 499–516. doi:10.1177/0962280206071644.

Maheswaran, R., Pearson, T., Smeeton, N. C., Beevers, S. D., Campbell, M. J., & Wolfe, C. D. (2012). Outdoor air pollution and incidence of ischemic and hemorrhagic stroke: a small-area level ecological study. *Stroke*, *43*, 22–7. doi:10.1161/STROKEAHA.110.610238.

Marmot, M. (2007). Achieving health equity: from root causes to fair outcomes. *The Lancet*, *370*, 1153–1163. doi:10.1016/S0140-6736(07)61385-3.

McCartney, G., Walsh, D., Whyte, B., & Collins, C. (2012). Has Scotland always been the 'sick man' of Europe? An observational study from 1855 to 2006. *European journal of public health*, *22*, 756–60. doi:10.1093/eurpub/ckr136.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. (2nd ed.). London: Chapman & Hall.

McMillan, N. J., Holland, D. M., Morara, M., & Feng, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics*, *21*, 48–65. doi:10.1002/env.984.

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

Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L., & Kaufman, J. D. (2007). Long-Term Exposure to Air Pollution and Incidence

of Cardiovascular Events in Women. *New England Journal of Medicine*, *356*, 447–458. doi:10.1056/NEJMoa054409.

Mills, I. C., Atkinson, R. W., Kang, S., Walton, H., & Anderson, H. R. (2015). Quantitative systematic review of the associations between short-term exposure to nitrogen dioxide and mortality and hospital admissions. *BMJ Open*, *5*, e006946–e006946. doi:10.1136/bmjopen-2014-006946.

Ministry of Public Health (1954). Mortality and Morbidity during the London Smog of December 1952. Her Majesty's Stationery Office volume 95.

Moolgavkar, S. H., Mcclellan, R. O., Dewanji, A., Turim, J., Georg Luebeck, E., & Edwards, M. (2013). Time-series analyses of air pollution and mortality in the united states: A subsampling approach. *Environmental Health Perspectives*, *121*. doi:10.1289/ehp.1104507.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*, 17–23.

Naess, Ø., Piro, F. N., Nafstad, P., Smith, G. D., & Leyland, A. H. (2007). Air pollution, social deprivation, and mortality: a multilevel cohort study. *Epidemiology*, *18*, 686–94. doi:10.1097/EDE.0b013e3181567d14.

National statistics. Scottish Index of Multiple Deprivation (2012). *A national statistics publication for Scotland*. Edinburgh: The Scottish Government.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*, 370–384. doi:10.2307/2344614.

Neuberger, M., Moshammer, H., & Rabczenko, D. (2013). Acute and Subacute Effects of Urban Air Pollution on Cardiopulmonary Emergencies and Mortality: Time Series Studies in Austrian Cities. *International Journal of Environmental Research and Public Health*, *10*, 4728–4751. doi:10.3390/ijerph10104728.

Omori, T., Fujimoto, G., Yoshimura, I., Nitta, H., & Ono, M. (2003). Effects of particulate matter on daily mortality in 13 Japanese cities. *Journal of Epidemiology / Japan Epidemiological Association*, *13*, 314–322.

O'Neill, M. S., Jerrett, M., Kawachi, I., Levy, J. I., Cohen, A. J., Gouveia, N., Wilkinson, P., Fletcher, T., Cifuentes, L., & Schwartz, J. (2003). Health, wealth, and air pollution: advancing theory and methods. *Environmental Health Perspectives*, *111*, 1861–70.

O'Neill, M. S., Loomis, D., & Borja-Aburto, V. H. (2004). Ozone, area social conditions, and mortality in Mexico City. *Environmental Research*, *94*, 234–42. doi:10.1016/j.envres.2003.07.002.

Ou, C.-Q., Hedley, A. J., Chung, R. Y., Thach, T.-Q., Chau, Y.-K., Chan, K.-P., Yang, L., Ho, S.-Y., Wong, C.-M., & Lam, T.-H. (2008). Socioeconomic disparities in air pollution-associated mortality. *Environmental research*, *107*, 237–44. doi:10.1016/j.envres.2008.02.002.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science : a review journal of the Institute of Mathematical Statistics*, *25*, 107–125. doi:10.1214/10-STS326.

Padilla, C. M., Deguen, S., Lalloue, B., Blanchard, O., Beaugard, C., Troude, F., Navier, D. Z., & Vieira, V. M. (2013). Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France. *The Science of the total environment*, *454-455*, 433–41. doi:10.1016/j.scitotenv.2013.03.027.

Pellow, D. (2000). Environmental Inequality Formation: Toward a Theory of Environmental Injustice. *American Behavioral Scientist*, *43*, 581–601. doi:10.1177/0002764200043004004.

Pirani, M., Gulliver, J., Fuller, G. W., & Blangiardo, M. (2014). Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science and Environmental Epidemiology*, *24*, 319–327. doi:10.1038/jes.2013.85.

Pleis, J. R., Lucas, J. W., & Ward, B. W. (2009). Summary health statistics for U.S. adults: National Health Interview Survey, 2008. *Vital and health statistics. Series 10, Data from the National Health Survey*, (pp. 1–157).

Poole, D., & Raftery, A. E. (2000). Inference for Deterministic Simulation Models: The Bayesian Melding Approach. *Journal of the American Statistical Association*, *95*, 1244. doi:10.2307/2669764.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. {*JAMA:*} {*The journal of the American Medical Association*}, *287*, 1132–1141.

Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association*, *56*, 709–42.

Pope III, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., & Heath, C. W. (1995). Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults. *American Journal of Respiratory and Critical Care Medicine*, *151*, 669–674. doi:10.1164/ajrccm/151.3_Pt_1.669.

Prescott, G. J. (2000). Investigation of factors which might indicate susceptibility to particulate air pollution. *Occupational and Environmental Medicine*, *57*, 53–57. doi:10.1136/oem.57.1.53.

Prescott, G. J., Cohen, G. R., Elton, R. A., Fowkes, F. G., & Agius, R. M. (1998). Urban air pollution and cardiopulmonary ill health: a 14.5 year time series study. *Occupational and environmental medicine*, *55*, 697–704.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: http://www.R-project.org/.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063.

Raïffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate School of Business Adminitration, Harvard University.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*, 1197–1206. doi:10.1111/j.1541-0420.2006.00617.x.

Richardson, E. A., Pearce, J., & Kingham, S. (2011). Is particulate air pollution associated with health and health inequalities in New Zealand? *Health & place*, *17*, 1137–43. doi:10.1016/j.healthplace.2011.05.007.

Richardson, S., Stücker, I., & Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International journal of epidemiology*, *16*, 111–20.

Roalfe, A. K., Holder, R. L., & Wilson, S. (2008). Standardisation of rates using logistic regression: a comparison with the direct method. *BMC Health Services Research*, *8*, 275. doi:10.1186/1472-6963-8-275.

Robert, C. P., & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. (1st ed.). Springer-Verlag New York. doi:10.1007/978-1-4419-1576-4.

Rosenlund, M., Picciotto, S., Forastiere, F., Stafoggia, M., & Perucci, C. A. (2008). Traffic-Related Air Pollution in Relation to Incidence and Prognosis of Coronary Heart Disease. *Epidemiology*, *19*, 121–128. doi:10.1097/EDE.0b013e31815c1921.

Rushworth, A., Lee, D., & Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and spatio-temporal epidemiology*, *10*, 29–38. doi:10.1016/j.sste.2014.05.001.

Sacks, J. D., Rappold, A. G., Allen Davis, J., Richardson, D. B., Waller, A. E., & Luben, T. J. (2014). Influence of urbanicity and county characteristics on the association between ozone and asthma emergency department visits in North Carolina. *Environmental Health Perspectives*, *122*, 506–512. doi:10.1289/ehp.1306940.

Sahu, S. K., Gelfand, A. E., & Holland, D. M. (2010). Fusing point and areal level spaceâ€“time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*, 77–103. doi:10.1111/j.1467-9876.2009.00685.x.

Samoli, E., Aga, E., Touloumi, G., Nisiotis, K., Forsberg, B., Lefranc, A., Pekkanen, J., Wojtyniak, B., Schindler, C., Niciu, E., Brunstein, R., Dodic Fikfak, M., Schwartz, J., & Katsouyanni, K. (2006). Short-term effects of nitrogen dioxide on mortality: an analysis within the APHEA project. *The European respiratory journal*, *27*, 1129–38. doi:10.1183/09031936.06.00143905.

Schofield, L., Walsh, D., Munoz-Arroyo, R., McCartney, G., Buchanan, D., Lawder, R., Armstrong, M., Dundas, R., & Leyland, A. H. (2016). Dying younger in Scotland: Trends in mortality and deprivation relative to England and Wales, 1981–2011. *Health & Place*, *40*, 106–115. doi:10.1016/j.healthplace.2016.05.007.

Schwartz, J., & Marcus, A. (1990). Mortality and air pollution in london: A time series analysis. *American Journal of Epidemiology; (USA)*, *131*, 185–194.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*, 461–464.

Scoggins, A., Kjellstrom, T., Fisher, G., Connor, J., & Gimson, N. (2004). Spatial analysis of annual air pollution exposure and mortality. *The Science of the total environment*, *321*, 71–85. doi:10.1016/j.scitotenv.2003.09.020.

Selvin, H. C. (1958). Durkheim's Suicide and Problems of Empirical Research. *American Journal of Sociology*, *63*, 607–619.

Shaddick, G., Lee, D., & Wakefield, J. (2013). Ecological bias in studies of the short-term effects of air pollution on health. *International Journal of Applied Earth Observation and Geoinformation*, *22*, 65–74. doi:10.1016/j.jag.2012.03.011.

Simpson, R., Williams, G., Petroeschevsky, A., Best, T., Morgan, G., Denison, L., Hinwood, A., Neville, G., & Neller, A. (2005). The short-term effects of air pollution on daily mortality in four Australian cities. *Australian and New Zealand journal of public health*, *29*, 205–12.

Smith, G. D., Bartley, M., & Blane, D. (1990). The Black report on socioeconomic inequalities in health 10 years on. *BMJ*, *301*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *64*, 583–616. doi:10.1111/1467-9868.00353.

Stewart, K. M. (1994). Air Pollution: Some Key Issues. *The Journal of the Royal Society for the Promotion of Health*, *114*, 86–90. doi:10.1177/146642409411400209.

Szpiro, A. a., & Paciorek, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, *24*, 501–517. doi:10.1002/env.2233.

Tao, Y., Huang, W., Huang, X., Zhong, L., Lu, S.-E., Li, Y., Dai, L., Zhang, Y., & Zhu, T. (2012). Estimated acute effects of ambient ozone and nitrogen dioxide on mortality in the Pearl River Delta of southern China. *Environmental health perspectives*, *120*, 393–8. doi:10.1289/ehp.1103715.

Terzano, C., Di Stefano, F., Conti, V., Graziani, E., & Petroianni, A. (2010). Air pollution ultrafine particles: toxicity beyond the lung. *European review for medical and pharmacological sciences*, *14*, 809–21.

Thach, T.-Q., Wong, C.-M., Chan, K.-P., Chau, Y.-K., Neil Thomas, G., Ou, C.-Q., Yang, L., Peiris, J. S., Lam, T.-H., & Hedley, A. J. (2010). Air pollutants and health outcomes: Assessment of confounding by influenza. *Atmospheric Environment*, *44*, 1437–1442. doi:10.1016/j.atmosenv.2010.01.036.

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

Tonne, C., Beevers, S., Armstrong, B., Kelly, F., & Wilkinson, P. (2008). Air pollution and mortality benefits of the London Congestion Charge: spatial and socioeconomic inequalities. *Occupational and environmental medicine*, *65*, 620–7. doi:10.1136/oem.2007.036533.

Tonne, C., Beevers, S., Kelly, F. J., Jarup, L., Wilkinson, P., & Armstrong, B. (2010). An approach for estimating the health effects of changes over time in air pollution: an illustration using cardio-respiratory hospital admissions in London. *Occupational and Environmental Medicine*, *67*, 422–7. doi:10.1136/oem.2009.048702.

Touloumi, G., Samoli, E., Quenel, P., Paldy, A., Anderson, R. H., Zmirou, D., Galan, I., Forsberg, B., Schindler, C., Schwartz, J., & Katsouyanni, K. (2005). Short-Term Effects of Air Pollution on Total and Cardiovascular Mortality. *Epidemiology*, *16*, 49–57. doi:10.1097/01.ede.0000142152.62400.13.

Townsend, P., Phillimore, P., & Beattie, A. (1988). *Health and Deprivation: Inequality and the North*. Croom Helm.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*, 2766–2772. doi:10.1890/07-0043.1.

Vicedo-Cabrera, A. M., Biggeri, A., Grisotto, L., Barbone, F., & Catelan, D. (2013). A Bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in Italy. *Geospatial Health*, *8*, 87–95.

Villeneuve, P. J., Burnett, R. T., Shi, Y., Krewski, D., Goldberg, M. S., Hertzman, C., Chen, Y., & Brook, J. (2003). A time-series study of air pollution, socioeconomic status, and mortality in Vancouver, Canada. *Journal of exposure analysis and environmental epidemiology*, *13*, 427–35. doi:10.1038/sj.jea.7500292.

Vinikoor-Imler, L. C., Davis, J. A., Meyer, R. E., & Luben, T. J. (2013). Early prenatal exposure to air pollution and its associations with birth defects in a state-wide birth cohort from North Carolina. *Birth defects research. Part A, Clinical and molecular teratology*, *97*, 696–701. doi:10.1002/bdra.23159.

Vinikoor-Imler, L. C., Davis, J. A., Meyer, R. E., Messer, L. C., & Luben, T. J. (2014). Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort. *Environmental Research*, *132*, 132–139. doi:10.1016/j.envres.2014.03.040.

Wakefield, J. (2004). Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*, 385–425. doi:10.1111/j.1467-985x.2004.02046_1.x.

Wakefield, J. (2008). Ecologic studies revisited. *Annual review of public health*, *29*, 75–90. doi:10.1146/annurev.publhealth.29.020907.090821.

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

Wakefield, J., & Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, *7*, 438–455. doi:10.1093/biostatistics/kxj017.

Walters, S., Phupinyokul, M., & Ayres, J. (1995). Hospital admission rates for asthma and respiratory disease in the West Midlands: their relationship to air pollution levels. *Thorax*, *50*, 948–54.

Wang, X. Y., Hu, W., & Tong, S. (2009). Long-term exposure to gaseous air pollutants and cardio-respiratory mortality in Brisbane, Australia. *Geospatial health*, *3*, 257–63.

Warren, J., Fuentes, M., Herring, A., & Langlois, P. (2012). Bayesian spatial-temporal model for cardiac congenital anomalies and ambient air pollution risk assessment. *Environmetrics*, *23*, 673–684. doi:10.1002/env.2174.

Warren, J. L., Herring, A. H., & Langlois, P. H. (2013). Periods of Pregnancy for Low Birth Weight, . *2013*, 7–9.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, *61*, 439–447. doi:10.1093/biomet/61.3.439.

Wilkinson, P., Elliott, P., Grundy, C., Shaddick, G., Thakrar, B., Walls, P., & Falconer, S. (1999). Case-control study of hospital admission with asthma in children aged 5-14 years: relation with road traffic in north west London. *Thorax*, *54*, 1070–4.

Willocks, L. J., Bhaskar, A., Ramsay, C. N., Lee, D., Brewster, D. H., Fischbacher, C. M., Chalmers, J., Morris, G., & Scott, E. M. (2012). Cardiovascular disease and air pollution in Scotland: no association or insufficient data and study design? *BMC Public Health*, *12*, 227. doi:10.1186/1471-2458-12-227.

Wong, C.-M., Ou, C.-Q., Chan, K.-P., Chau, Y.-K., Thach, T.-Q., Yang, L., Chung, R. Y.-N., Thomas, G. N., Peiris, J. S. M., Wong, T.-W., Hedley, A. J., & Lam, T.-H. (2008). The effects of air pollution on mortality in socially deprived urban areas in Hong Kong, China. *Environmental Health Perspectives*, *116*, 1189–94. doi:10.1289/ehp.10850.

World Health Organisation (1975). *Manual of the international Statistical Classification of Diseases, Injuries and Cause of Death, 9th revision.*. URL: http://apps.who.int/iris/bitstream/10665/40492/1/9241540044_eng_v1_p1.pdf.

World Health Organisation (1994). *Manual of the International Statistical Classification of Diseases and Related Health Problems, 10th revision.*. URL: http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf?ua=1.

World Health Organisation (2006). Air quality guidelines. Global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide.

World Health Organisation (2014). Ambient (outdoor) air quality and health: Fact Sheet 313. URL: http://www.who.int/mediacentre/factsheets/fs313/en/.

Yap, C., Beverland, I. J., Heal, M. R., Cohen, G. R., Robertson, C., Henderson, D. E. J., Ferguson, N. S., Hart, C. L., Morris, G., & Agius, R. M. (2012). Association between long-term exposure to air pollution and specific causes of mortality in Scotland. *Occupational and Environmental Medicine*, *69*, 916–24. doi:10.1136/oemed-2011-100600.

Yi, O., Kim, H., & Ha, E. (2010). Does area level socioeconomic status modify the effects of PM(10) on preterm delivery? *Environmental research*, *110*, 55–61. doi:10.1016/j.envres.2009.10.004.

Zhu, L., Carlin, B. P., & Gelfand, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, *14*, 537–557. doi:10.1002/env.614.

Zidek, J. V., Shaddick, G., & Taylor, C. G. (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *The Annals of Applied Statistics*, *8*, 1640–1670. doi:10.1214/14-AOAS745.