UNIVERSITY OF GLASGOW

Quantification of Air Quality in Space and Time and Its Effects on Health

by

Guowen Huang

A thesis submitted to the University of Glasgow for the degree of Doctor of Philosophy

in

Statistics

October 2016

Declaration of Authorship

I, Guowen Huang, declare that this thesis titled, 'Quantification of Air Quality in Space and Time and Its Effects on Health' and the work presented in it are my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

The work presented in Chapter 4 has been published in Spatial and Spatio-temporal Epidemiology with the title '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' (September 2015). This work has also been presented at the GEOMED conference hold in Florence, Italy, 2015. The work presented in Chapter 5 and 6 have been presented at the 26th Annual Conference of The International Environmetrics Society, and a manuscript based on them is currently in preparation.

Signed:

Date:

"WHERE THE LAND ENDS AND THE SEA BEGINS."

Carte de Mons

Abstract

The long-term adverse effects on health associated with air pollution exposure can be estimated using either cohort or spatio-temporal ecological designs. In a cohort study, the health status of a cohort of people are assessed periodically over a number of years, and then related to estimated ambient pollution concentrations in the cities in which they live. However, such cohort studies are expensive and time consuming to implement, due to the long-term follow up required for the cohort. Therefore, spatio-temporal ecological studies are also being used to estimate the long-term health effects of air pollution as they are easy to implement due to the routine availability of the required data. Spatio-temporal ecological studies estimate the health impact of air pollution by utilising geographical and temporal contrasts in air pollution and disease risk across n contiguous small-areas, such as census tracts or electoral wards, for multiple time periods. The disease data are counts of the numbers of disease cases occurring in each areal unit and time period, and thus Poisson log-linear models are typically used for the analysis. The linear predictor includes pollutant concentrations and known confounders such as socio-economic deprivation. However, as the disease data typically contain residual spatial or spatio-temporal autocorrelation after the covariate effects have been accounted for, these known covariates are augmented by a set of random effects. One key problem in these studies is estimating spatially representative pollution concentrations in each areal which are typically estimated by applying Kriging to data from a sparse monitoring network, or by computing averages over modelled concentrations (grid level) from an atmospheric dispersion model.

The aim of this thesis is to investigate the health effects of long-term exposure to Nitrogen Dioxide (NO_2) and Particular matter (PM_{10}) in mainland Scotland, UK. In order to have an initial impression about the air pollution health effects in mainland Scotland, chapter 3 presents a standard epidemiological study using a benchmark method. The remaining main chapters (4, 5, 6) cover the main methodological focus in this thesis which has been threefold: (i) how to better estimate pollution by developing a multivariate spatio-temporal fusion model that relates monitored and modelled pollution data over space, time and pollutant; (ii) how to simultaneously estimate the joint effects of multiple pollutants; and (iii) how to allow for the uncertainty in the estimated pollution concentrations when estimating their health effects.

Specifically, chapters 4 and 5 are developed to achieve (i), while chapter 6 focuses on (ii) and (iii). In chapter 4, I propose an integrated model for estimating the long-term health effects of NO_2, that fuses modelled and measured pollution data to provide improved predictions of areal level pollution concentrations and hence health effects. The air pollution fusion model proposed is a Bayesian space-time linear regression model for relating the measured concentrations to the modelled concentrations for a single pollutant, whilst allowing for additional covariate information such as site type (e.g. roadside, rural, etc) and temperature. However, it is known that some pollutants might be correlated because they may be generated by common processes or be driven by similar factors such as meteorology. The correlation between pollutants can help to predict one pollutant by borrowing strength from the others. Therefore, in chapter 5, I propose a multi-pollutant model which is a multivariate spatio-temporal fusion model that extends the single pollutant model in chapter 4, which relates monitored and modelled pollution data over space, time and pollutant to predict pollution across mainland Scotland.

Considering that we are exposed to multiple pollutants simultaneously because the air we breathe contains a complex mixture of particle and gas phase pollutants, the health effects of exposure to multiple pollutants have been investigated in chapter 6. Therefore, this is a natural extension to the single pollutant health effects in chapter 4. Given NO_2 and PM_{10} are highly correlated (multicollinearity issue) in my data, I first propose a temporally-varying linear model to regress one pollutant (e.g. NO_2) against another (e.g. PM_{10}) and then use the residuals in the disease model as well as PM_{10}, thus investigating the health effects of exposure to both pollutants simultaneously. Another issue considered in chapter 6 is to allow for the uncertainty in the estimated pollution concentrations when estimating their health effects. There are in total four approaches being developed to adjust the exposure uncertainty.

Finally, chapter 7 summarises the work contained within this thesis and discusses the implications for future research.

Acknowledgements

Firstly, I would like to thank my supervisors Dr. Duncan Lee and Prof. Marian Scott for their invaluable support and guidance throughout all of the stages of my PhD research. I did not specialise in statistics before I arrived in Glasgow University, therefore I am extremely grateful for their encouragement and patience, without which I would not have been able to develop this thesis. In addition, I gratefully acknowledge the funding from the China Scholarships Council (CSC) that allowed me to undertake this work.

Thank you to everyone in the Department of Statistics, especially all my office mates in room 120 (Kelly, Ruth, Daniel, Amira, Mengyi, Craig, Aisyah, Sami, Emmanuel) for providing such diverse and helpful environments, and Dr. Ludger Evers and Dr. Claire Miller who helped me a lot in the annual mini vivas. To the stats badminton team (Duncan, Ludger, Francesca, Craig, Vinny) and the APTS team (Francesca, Tushar, Elaine, Carol, Rebecca, Mengyi, Basile, Alasdair), thank you so much for bringing me lots of enjoyable time and wonderful memories.

Big thanks are due to Glasgow University Student Learning Service (SLS), where I worked as a stats tutor for two years and also learned a lot through solving statistical problems for hundreds of students. To Shazia, Carol, Mags, thank you so much for helping me a lot during my stay in SLS.

Finally, to my family, thank you for support and encouragement always. To my Chinese friends in Glasgow, thank you all for supporting me a lot during my PhD, for beers, badminton, travels, dinners, parties, and karaoke.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements v

List of Figures ix

List of Tables xi

1 Introduction 1
- 1.1 Pollutants and their potential health effects . . . 2
- 1.2 Methods for examing air pollution health effects . . . 4
- 1.3 Research region . . . 6
- 1.4 Thesis structure . . . 9

2 Statistical background 10
- 2.1 The normal linear model . . . 10
  - 2.1.1 Model fitting . . . 12
  - 2.1.2 Model checking . . . 13
  - 2.1.3 Added variable plot . . . 14
- 2.2 Generalised linear models . . . 15
- 2.3 Spatial methods . . . 16
  - 2.3.1 Geostatistical data . . . 16
  - 2.3.2 Areal data . . . 25
    - 2.3.2.1 Spatial association in areal data . . . 26
    - 2.3.2.2 Spatial modelling for areal data . . . 26
    - 2.3.2.3 Spatio-temporal modelling for areal data . . . 29
- 2.4 Bayesian modelling . . . 30
  - 2.4.1 Bayes' theorem . . . 30
  - 2.4.2 The first stage of a Bayesian model: probability model . . . 31
  - 2.4.3 The second stage of the Bayesian model: prior . . . 31
  - 2.4.4 Using the data to update the prior: posterior distribution . . . 32
  - 2.4.5 Inference . . . 34
    - 2.4.5.1 Markov chain Monte Carlo simulation . . . 34

3 The impact of air pollution on health 39

3.1 Introduction . . . 39
3.2 Data description . . . 41
3.3 Exploratory analysis . . . 43
3.4 Methods . . . 45
3.5 Results . . . 47
3.5.1 The effect of NO_2 on health . . . 47
3.5.2 The effect of PM_{10} on health . . . 50
3.6 Discussion . . . 52

4 Estimating the long-term health effects of air pollution by fusing modelled and measured pollution data 54
4.1 Introduction . . . 54
4.2 Background . . . 57
4.2.1 Data description . . . 57
4.2.2 Exploratory analysis . . . 63
4.2.3 Spatio-temporal pollution modelling . . . 65
4.3 Methodology . . . 65
4.3.1 Stage 1 - air pollution model . . . 66
4.3.2 Stage 2 - disease model . . . 80
4.4 Results . . . 82
4.4.1 Pollution model validation . . . 82
4.4.2 Pollution model prediction . . . 83
4.4.3 Disease model results . . . 85
4.4.4 Sensitivity analysis . . . 87
4.5 Discussion . . . 88

5 Multi-pollutant concentrations prediction 91
5.1 Introduction . . . 91
5.2 Data description . . . 92
5.3 Methodology . . . 95
5.3.1 Modelling . . . 96
5.3.2 Computation of the posterior distribution . . . 98
5.3.3 Dealing with missing data . . . 102
5.4 Simulation study . . . 103
5.4.1 Data generation . . . 103
5.4.2 Simulation method . . . 104
5.4.3 Simulation results . . . 105
5.5 Validation study . . . 112
5.6 Application . . . 113
5.6.1 Model fitting . . . 114
5.6.2 Model prediction . . . 116
5.7 Conclusion . . . 118

6 Health Effects of Exposure to Multiple Pollutants 120
6.1 Introduction . . . 120
6.2 Data description . . . 124
6.3 Methodology . . . 125

6.3.1 Single pollutant disease model . . . 125
6.3.2 Co-pollutant disease model . . . 125
6.3.3 Multi-pollutant disease model . . . 126
6.3.4 Dealing with exposure uncertainty . . . 130
6.4 Results . . . 136
6.4.1 Single pollutant health effects . . . 137
6.4.2 Co-pollutant health effects . . . 138
6.4.3 Multi-pollutant health effects . . . 138
6.4.4 Health effects with consideration of exposure uncertainty . . . 142
6.4.4.1 Results from approach 1 . . . 143
6.4.4.2 Results from approach 2 . . . 145
6.4.4.3 Results from approach 3 . . . 145
6.4.4.4 Results from approach 4 . . . 148
6.5 Discussion and conclusion . . . 151

7 Conclusion **154**
7.1 Initial impression of the air pollution health effects . . . 155
7.2 Improved air pollution predictions - single-pollutant model . . . 156
7.3 Improved air pollution predictions - multi-pollutant model . . . 157
7.4 Single pollutant health effects . . . 157
7.5 Multiple pollutants health effects . . . 158
7.6 Dealing with exposure uncertainty . . . 159
7.7 Key themes . . . 160
7.8 Discussions and future work . . . 161

Bibliography **165**

List of Figures

2.1 Exponential covariance function and associated semi-variogram. 20

3.1 The distributions of the observed and expected admissions in 2011 and their standardized incidence ratio: top left is the observed hospital admissions, top right is the expected hospital admissions, two figures in the bottom are the corresponding SIR. 42

3.2 The spatial distributions of NO_2 and PM_{10} for 2010 in Scotland (unit: μgm^{-3}). Top left is based on using the mean gridded NO_2 concentrations in each IG, top right is based on using the max gridded NO_2 concentrations in each IG, bottom left is based on using the mean gridded PM_{10} concentrations in each IG, bottom right is based on using the max gridded PM_{10} concentrations in each IG. 44

3.3 Scatterplots of log respiratory disease SIR (log(SIR)) against Job Seekers Allowance (JSA) and log of median property price (logprice). 45

3.4 Model residuals from fitting a Poisson model for maximum NO_2 with the Leroux CAR prior. 48

3.5 Model residuals from fitting a Poisson model for maximum PM_{10} with the Leroux CAR prior. 51

4.1 Summary of the data. Top left is the SIR for respiratory disease in Scotland in 2011, top right is the modelled annual average NO_2 concentrations in 2010 (μgm^{-3}), bottom left is the locations of the measured NO_2 data (▲ for monitoring sites, + for tube sites), and bottom right shows Scotland partitioned into urban (black) and rural areas (grey). 58

4.2 Summary of SIR. Top left and right are the SIR for respiratory disease in Scotland in 2007 and 2008, respectively, while the bottom left and right are for 2009 and 2010, respectively. 59

4.3 Summary of modelled NO_2. Top left and right are the modelled NO_2 for respiratory disease in Scotland in 2006 and 2007, respectively, while the bottom left and right are for 2008 and 2009, respectively. 61

4.4 The empirical semi-variogram of the residuals from the geostatistical model for 2010 (circles), with 95% Monte Carlo simulation envelopes (dashed lines). 64

4.5 Spatially aggregated predicted NO_2 concentrations (μgm^{-3}) from Model 1A for 2010. The left panel shows the spatial mean concentration over each IG, while the right panel shows the spatial maximum concentration over each IG. 85

4.6 Comparison between DEFRA and predicted NO_2 concentrations (μgm^{-3}) from Model 1A for 2010. The left panel is a scatterplot of all model-predicted versus DEFRA values while the right panel shows a spatial map of the differences between them. 86

5.1 Summary of the data. Top left is a map of the monitoring sites for both NO_2 and PM_{10} in 2010 (▲: common sites; red +: sites with only NO_2; blue •: sites with only PM_{10}), top right is the modelled annual average PM_{10} concentrations in 2010 (μgm^{-3}), bottom left are scatter plots between measured NO_2 and measured PM_{10}, modelled NO_2 and modelled PM_{10} for common sites, bottom right are scatter plots between measured and modelled data for common sites. 94
5.2 The results of the 30,000 McMC simulations for the overall mean of the regression parameters. 115
5.3 The predicted NO_2 and PM_{10} for 2010 from multi-pollutant model based on 1km resolution (unit: μgm^{-3}). Top left is the predicted NO_2 and top right is its predicted standard deviation, bottom left is the predicted PM_{10} and bottom right is its predicted standard deviation. 117
5.4 The predicted NO_2 and PM_{10} for 2010 from multi-pollutant model based on IG scale (unit: μgm^{-3}). Top left is based on using the max gridded NO_2 concentrations in each IG, top right is based on using the mean gridded NO_2 concentrations in each IG, bottom left is based on using the max gridded PM_{10} concentrations in each IG, bottom right is based on using the mean gridded PM_{10} concentrations in each IG. 119

6.1 Four approaches to adjust for exposure uncertainty: X_{1kt} could be NO_2 or PM_{10}. 130
6.2 ACF plots of X^i_{1kt} for a randomly selected (k,t). 133
6.3 Normal qq plots of X^i_{1kt} for a randomly selected (k,t). 134
6.4 Scatter plots of the variance of X_{1kt} against $\bar{X}_{1kt}$ and $(\bar{X}_{1kt})^2$ for maximum NO_2. 135
6.5 Model residuals from regressing spatial maximum PM_{10} against spatial maximum NO_2. 140
6.6 Standard deviations (sd: μgm^{-3}) of the exposure within IGs: with the red lines are the sd of the exposure across IGs while exposure error is not considered. 144
6.7 McMC trace plot for σ_1^2, σ_2^2 and a randomly selected $(X_{1kt},\ \hat{\epsilon}_{kt})$ from approach 3 by using spatial maximum of PM_{10}. 147
6.8 McMC trace plot for σ_1^2, σ_2^2 and a randomly selected $(X_{1kt},\ \hat{\epsilon}_{kt})$ from approach 4 by using spatial maximum of PM_{10}. 149

List of Tables

2.1 Commonly used conjugate priors. 32

3.1 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for mean NO_2 with a Leroux random effect. 49
3.2 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for max NO_2 with a Leroux random effect. 49
3.3 Relative risk for a $6.84\mu gm^{-3}$ increase of NO_2 for eight models based on mean NO_2 data in each IG. 49
3.4 Relative risk for a $6.84\mu gm^{-3}$ increase of NO_2 for eight models based on maximum NO_2 data in each IG. 50
3.5 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for mean PM_{10} with a Leroux random effect. 51
3.6 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for max PM_{10} with a Leroux random effect. 51
3.7 Relative risk for a $1.872\mu gm^{-3}$ increase of PM_{10} from the eight models based on mean PM_{10} data in each IG. 52
3.8 Relative risk for a $1.872\mu gm^{-3}$ increase of PM_{10} for eight models based on maximum PM_{10} data in each IG. 52

4.1 Summary of the measured NO_2 data by site type and year: the numbers within the round brackets represent the number of sites in the form (automatic monitors, diffusion tubes), while those within square brackets indicate their corresponding mean concentrations (μgm^{-3}). 60
4.2 Scottish Government 8 fold Urban Rural Classification 63
4.3 Simplifications of the general model (4.3). 71
4.4 Bias, root mean square prediction error and coverage probabilities from a 10 fold cross validation excercise for the models proposed in this chapter, the autoregressive Gaussian process model (**SGH**) and using only the DEFRA concentrations. 84
4.5 Posterior means for the regression parameters from Model 1A and the Gaussian process model **SGH**. The five columns $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5)$ are the yearly regression parameter estimates from Model 1A, whlie Model **SGH** has constant regression parameters over time (final column). 84

4.6 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the disease model (4.21) with four different pollution metrics. Model I and II correspond to the spatial mean and maximum of the DEFRA concentrations, while Models III and IV relate to the spatial mean and maximum of the predicted pollution concentrations from Model 1A. The regression parameters are presented as relative risks for a standard deviation increase in each covariates value, which is NO_2 6.84 μgm^{-3}, Logprice 0.38, JSA 2.35. 87
4.7 Relative risk of NO_2 against various basis dimensions of the space smoothness. 88

5.1 Numbers of monitoring sites measuring NO_2 and PM_{10}. 93
5.2 Summary of the monitoring data by site type and year: the numbers within the round brackets represent the number of sites in the form (NO_2, PM_{10}), while those within square brackets indicate their corresponding mean concentrations (μgm^{-3}). 93
5.3 Simulation settings of regression parameters for Model (5.1) 104
5.4 The bias of each parameter from the simulation study of Model (5.1). . . 106
5.5 The RMSE of each parameter from the simulation study of Model (5.1). . 107
5.6 The CI coverage (%) for each parameter from the simulation study of Model (5.1). 108
5.7 Bias, RMSE and 95% coverage (the same order in each bracket) for the covariance in the simulation study of Model (5.1) for more levels of correlation between NO_2 and PM_{10}. 109
5.8 Comparison of bias, RMSE and 95% coverage (the same order in each bracket) for fixed coefficients between the simulation study of Model (5.1) (round brackets) and single-pollutant model (4.3) (square brackets) for more levels of correlation between NO_2 and PM_{10}. 110
5.9 Bias, root mean square prediction error and coverage probabilities from a leave one out cross validation excercise for the single pollutant model (4.3) and the multipollutant model (5.1), based on the the common sites in 2010. 112
5.10 Results from the single pollutant model (4.3) and the multipollutant model (5.1) for 2010. 113
5.11 Posterior means for the regression parameters from multi-pollutant model. 114

6.1 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the single pollutant disease model. 137
6.2 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the co-pollutant disease model. 139
6.3 Model summaries from model (6.2) (unit for residuals: μgm^{-3}) 140
6.4 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model. 141
6.5 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the PCA disease model (the relative risks are based on the same increasing units with those in Table 6.4). 142

6.6 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 1. 143
6.7 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the single pollutant disease model using approach 2. 146
6.8 Posterior mean and 95% credible intervals for σ_1^2 (variance of X_{1kt} within an IG) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) from model (6.7). 146
6.9 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 3. 148
6.10 Posterior mean and 95% credible intervals for σ_1^2 (slope between var(X_{1kt}) and X_{1kt}^2) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) from model (6.8). 149
6.11 Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 4. 150

Chapter 1

Introduction

A recent World Health Organisation report estimates that outdoor air pollution was responsible for the premature deaths of 3.7 million people under the age of 60 in 2012 (World Health Organisation [142]), which indicates that the air we breathe has a major impact on our health and the environment, and air pollution is a global health problem. The health impact of air pollution exposure has been widely recognised since the 1950's, as a result of the London smog in December 1952, which is estimated to have resulted in more than 3,000 excess deaths compared with previous years (Bell and Davis [9]). To combat this problem, the Clean Air Act was introduced by the UK government in 1956, which was an important milestone in the development of a legal framework to protect the environment. The act introduced a number of measures to reduce air pollution, especially by introducing "smoke control areas" in some towns and cities in which only smokeless fuels could be burned. Since then additional legislation has been introduced, including further clean air acts in 1968 and 1993, the UK Air Quality Strategy in 1997, 2000 and 2007 (Department for Environment and Food And Rural Affairs [32]), and the 2008 ambient air quality directive (2008/50/EC). The Air Quality Strategy established the framework for air quality improvements across the UK and set out the Air Quality Standards and Objectives which has been set to benchmark air quality in terms of protecting human health and the environment. For example, daily mean particulate matter concentrations (measured as PM_{10}) must not exceed $50\mu gm^{-3}$ more than 35 times a year. The 2008 ambient air quality directive (2008/50/EC) sets legally binding limits for concentrations in ambient (outdoor) air of major air pollutants that are known to have a significant impact on human health including particulate matter (PM_{10} and

$PM_{2.5}$) and nitrogen dioxide (NO_2). As a result, pollution levels today are continuously measured by a network of monitors, which record ambient (background) concentrations at both road-side and background environments. Numerous pollutants are measured by this network, including Ozone (O_3), Carbon monoxide (CO), Sulphur dioxide (SO_2), NO_2, PM_{10} and $PM_{2.5}$.

1.1 Pollutants and their potential health effects

The adverse effects of air pollution on health have been widely investigated all over the world, and various pollutants which have adverse effects on health have been reported, including O_3, CO, SO_2, NO_2, PM_{10} and $PM_{2.5}$.

O_3 is not emitted directly from any human-made source. It arises from chemical reactions between various air pollutants, primarily NO_X (NO and NO_2) and Volatile Organic Compounds (VOCs), initiated by strong sunlight. Since O_3 is a highly reactive substance, any adverse health effects will be found essentially at the sites of initial contact: the respiratory tract (nose, throat and airways), the lungs, and at higher concentrations, the eyes (Health and Safety Executive [55]). Very high levels can damage airways leading to inflammatory reactions. Ozone reduces lung function and increases incidence of respiratory symptoms, respiratory hospital admissions and mortality (Department for Environment and Food And Rural Affairs [32]).

CO is formed from incomplete combustion of carboncontaining fuels. Road transport is the largest source of CO, and residential and industrial combustion also make significant contributions. Exposure to CO can reduce the oxygen-carrying capacity of the blood, which is risky for those people with a reduced capacity for pumping oxygenated blood to the heart due to several types of heart disease. Exposure to CO can cause them to experience myocardial ischemia (reduced oxygen to the heart), often accompanied by chest pain (angina), when exercising or under increased stress (United States Environmental Protection Agency: *https://www3.epa.gov/*).

SO_2 emissions are usually dominated by combustion of fuels containing sulphur, such as coal and heavy oils by power stations and refineries. For people suffering from asthma and chronic lung disease already, SO_2 can likely cause constriction of the airways of the lung. Exposure to high levels of SO_2 can have potential damage to ecosystems,

including degradation of chlorophyll, reduced photosynthesis, raised respiration rates and changes in protein metabolism (Department for Environment and Food And Rural Affairs [32]). Recent research has also reported the adverse effects of SO_2. For example, Brown et al. [18] reported that the odds of admission for all respiratory diseases were statistically significantly greater in the SO_2 episode week than in the control week (odds ratio 1.40, with a 95% confidence interval (1.00, 1.94)). Elliott et al. [41] found significant associations between BS and SO_2 concentrations and mortality. Chen et al. [27] found that an increase of $10\mu gm^{-3}$ in the two-day moving average SO_2 concentration was associated with 0.75% (95% posterior interval (PI), 0.47% to 1.02%), 0.83% (0.95% PI, 0.47% to 1.19%) and 1.25% (95% PI, 0.78% to 1.73%) increase of total, cardiovascular and respiratory mortality, respectively.

NO_2 and nitric oxide (NO) are both oxides of nitrogen and together are referred to as NO_X which are produced by combustion processes in air. Road transport is the main source, followed by the electricity supply industry and other industrial and commercial sectors. Exposure to high levels of NO_2 can cause inflammation of the airways. Long term exposure may affect lung function and respiratory symptoms. NO_2 also enhances the response to allergens in sensitive individuals and contributes to the formation of secondary particles and ground level O_3, both of which are associated with ill-health effects. Chauhan et al. [26] reported that NO_2 increased the susceptibility to respiratory infections, while Lee et al. [76] showed that long-term exposure (over 3 years) to NO_2 was significantly associated with respiratory hospital admissions in Edinburgh and Glasgow.

Particulate Matter (PM_{10} and $PM_{2.5}$) is generally categorised on the basis of the size of the particles (for example $PM_{2.5}$ is particles with a diameter of less than 2.5 micrometres). Particulate Matter is made up of a wide range of materials and arises from both human-made (such as stationary fuel combustion and transport) and natural sources (such as sea spray and Saharan dust). Concentrations of Particulate Matter comprises primary particles emitted directly into the atmosphere from combustion sources and secondary particles formed by chemical reactions in the air. Exposure to Particulate Matter is associated with respiratory and cardiovascular illness and mortality as well as other ill-health effects. These associations are believed to be causal because PM_{10} roughly equates to the mass of particles less than 10 micrometres in diameter that are likely to be inhaled into the thoracic region of the respiratory tract. Lee et al. [76] showed that long-term exposure (over 3 years) to PM_{10} was significantly associated with respiratory

hospital admissions in Edinburgh and Glasgow, while Lee [74] reported that a $1.7\mu gm^{-3}$ increase in PM_{10} concentrations was associated with 6.6% additional hospital admissions due to respiratory diseases across Scotland. Recent reviews (Committee on the Medical Effects of Air Pollutants [29]) have suggested exposure to a finer fraction of particles ($PM_{2.5}$) give a stronger association with the observed ill-health effects because they can travel deeper into lungs. US Environmental Protection Agency [133] also reported an increase of 1% (range 0.4% to 1.8%) in annual all-cause deaths for a $1\mu gm^{-3}$ increase in the annual average of $PM_{2.5}$ exposure in the United States.

1.2 Methods for examing air pollution health effects

According to the existing literature in epidemiological studies, the adverse effects on health associated with air pollution exposure can be considered in two ways, the short-term and long-term effects. These refer to studies investigating associations of health effects with variations in ambient pollution concentrations averaged over a short-time period (often daily averages) or long-term differences in concentrations (often annual averages).

For short-term effects, daily counts of disease cases are regressed against air pollution concentrations on the preceding few days, utilising an ecological (at the population level) time series design. A number of existing studies in epidemiological studies were focused on the adverse effects on health associated with short-term exposure to air pollution, with examples including Neukirch et al. [93], Kontos et al. [69], Katsouyanni et al. [65], Dominici et al. [38], Desqueyroux et al. [34], Gilmour et al. [47], Goldberg and Burnett [49], Wong et al. [141] and Lee et al. [79].

In contrast, the long-term effects of air pollution can be estimated using either cohort or spatio-temporal ecological designs. In a cohort study, the health status of a cohort of people are assessed periodically over a number of years, and then related to estimated ambient pollution concentrations in the cities in which they live, with examples including American cohort studies - the Six Cities study (Dockery et al. [37], Laden et al. [71]), the American Cancer Society (ACS) study (Pope et al. [102]) and several European cohort studies reporting long term effects of air pollution on mortality(Hoek et al. [60], Hoek

et al. [59], Nafstad et al. [92], Raaschou-Nielsen et al. [106], Carey et al. [23], Cesaroni et al. [25], Beelen et al. [7], Stockfelt et al. [122]).

However, such cohort studies are expensive and time consuming to implement, due to the long-term follow up required for the cohort. Therefore spatio-temporal ecological studies are also being used to estimate the long-term health effects of air pollution as they are easy to implement due to the routine availability of the required data. Examples of such studies in a purely spatial context include Jerrett et al. [63], Maheswaran et al. [85], Maheswaran et al. [84], Barceló et al. [5], Lee et al. [76], Young et al. [143], Haining et al. [53], Lee [74], while spatio-temporal designs include Elliott et al. [41], Janes et al. [62], Greven et al. [52], Lawson et al. [72], Rushworth et al. [109].

Spatio-temporal ecological studies estimate the health impact of air pollution by utilising geographical and temporal contrasts in air pollution and disease risk across n contiguous small-areas, such as census tracts or electoral wards, for multiple time periods. The disease data are counts of the numbers of disease cases occurring in each areal unit and time period, and thus Poisson log-linear models are typically used for the analysis. The linear predictor includes pollutant concentrations and known confounders such as socio-economic deprivation. However, the disease data typically contain residual spatial or spatio-temporal autocorrelation after the covariate effects have been accounted for, which is caused by numerous factors, including unmeasured confounding, neighbourhood effects (where subjects behaviour is influenced by neighbouring subjects) and grouping effects (where subjects choose to be close to similar subjects) (Rushworth et al. [109]). Therefore, these known covariates are augmented by a set of random effects which are commonly modelled by the class of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model. Spatial correlation between the random effects is determined by a binary $n \times n$ neighbourhood matrix $\boldsymbol{W}$. Based on this neighbourhood matrix, the most common models for the random effects include intrinsic autoregressive (IAR) (Besag et al. [15]), convolution or BYM model (Besag et al. [15]), as well as those proposed by Cressie [31] and Leroux et al. [80]. These CAR models differ by holding different assumptions about how the random effects depend on each other across space, more details can be seen in chapter 2. While the study is a spatio-temporal design, Rushworth et al. [109] proposed a CAR model which allows for the residual spatio-temporal autocorrelation.

One key problem in these studies is estimating spatially representative pollution concentrations, using either measured data from a sparse network of monitors or modelled concentrations on a regular grid from an atmospheric dispersion model, such as those produced by AEA [1]. The latter provide complete spatial coverage of the study region but are known to contain biases (Berrocal et al. [11]). However, the monitored (point locations) and modelled (grid boxes) pollution data are spatially misaligned with the disease data (irregularly shaped areal units), and the problem about dealing with multiple data sources on different scales is often referred to as the *change of support problem* (Gelfand et al. [44], Gotway and Young [50]). There are a few epidemiological studies which use monitored pollution data alone to estimate spatially representative pollution concentrations, where geostatistical Kriging has been used to spatially align the monitored pollution data to the disease counts (Elliott et al. [41], Janes et al. [62]). In contrast, some studies use modelled concentrations alone to estimate spatially representative pollution concentrations, where simple averaging is used to correct the spatial misalignment of the modelled concentrations (Maheswaran et al. [84], Lee et al. [76], Warren et al. [136]). Recently, Vinikoor-Imler et al. [134], Vinikoor-Imler et al. [135], Sacks et al. [111] and Warren et al. [137] have estimated pollution using both monitored and modelled pollution data, by utilizing the fusion approaches proposed by Fuentes and Raftery [42], Berrocal et al. [11] or McMillan et al. [87].

1.3 Research region

The United Kingdom (UK) is a sovereign state in Europe, which lies off the north-western coast of the European mainland and includes the island of Great Britain, the north-eastern part of the island of Ireland and many smaller islands. The UK has a population of about 64 million people, which consists of four countries: England (53.8 million), Scotland (5.3 million), Wales (3.1 million), and Northern Ireland (1.8 million).

In the UK, the number of annual deaths are around half a million which is roughly 0.1% of the population. Three main leading causes of death are cancer (International Classification of Disease version 10 codes C00-D48), cardiovascular disease (International Classification of Disease version 10 codes I00-I99) and respiratory disease (International Classification of Disease version 10 codes J00-J99). According to Townsend et al. [131], the total number of deaths in the UK in 2014 is 570,341, among which

167,582 (29.4%) deaths are caused by cancer, 154,639 (27.1%) by cardiovascular disease and 75,282 (13.2%) by respiratory disease. The distribution of the deaths across the four countries in the UK is: 468,875 deaths from England, 54,239 from Scotland, 31,439 from Wales and 14,678 from Northern Ireland.

The health effects of air pollution are still significant. Evidence from the Department for Environment and Food And Rural Affairs [33] suggests that the effects of NO_2 on mortality are equivalent to 23,500 deaths annually in the UK. The impact of exposure to particulate matter is estimated to have an effect on mortality equivalent to nearly 29,000 deaths in the UK (Committee on the Medical Effects of Air Pollutants [29]).

The main aim of this thesis is to investigate the health effects of long-term exposure to air pollution (NO_2 and PM_{10}) in mainland Scotland. There have been few previous related epidemiological studies in Scotland, for example, only Prescott et al. [104], Carder et al. [22] and Willocks et al. [139] have investigated the association between short-term exposure to air pollution and ill health, while only Lee et al. [76] and Lee [74] have attempted to quantify the long-term effects of exposure using an ecological spatio-temporal design.

The Air Quality in Scotland website (*http://www.scottishairquality.co.uk/*) provides ample information about air pollution in Scotland, such as a Daily Air Quality Index (DAQI), an interactive map which can be used to explore different Scottish air quality monitoring sites, and a database containing tables of measured concentration data and statistics from the air quality monitoring sites.

The DAQI uses the index and banding system approved by the Committee on Medical Effects of Air Pollution (COMEAP) which uses 1-10 index divided into four bands to provide more detail about air pollution levels in a simple way, 1-3 (Low), 4-6 (Moderate), 7-9 (High), 10 (Very High). The overall air pollution index for a site or region is calculated from the highest concentration of five pollutants: NO_2, SO_2, O_3, $PM_{2.5}$ and PM_{10}.

There are in total 91 monitoring sites across Scotland, which consist of the automatic networks that measure various pollutants, including NO_2, CO, SO_2, O_3, $PM_{2.5}$ and PM_{10}. Monitoring sites can be classified according to the types of environment in which they are located, in order to permit more meaningful evaluation of data, such as rural,

urban background and roadside. The automatic networks produce hourly pollutant concentrations, with the data going back as far as 1986 at some sites. A range of simple statistics are routinely calculated by the database for the automatic monitoring data, including: daily mean, maximum and minimum values for all pollutants, 8-hour running mean values for O_3 and CO, daily maximum 8-hour running means for O_3, running 24-hour means for particulate matter. In addition, monthly and annually statistics are also provided.

In this thesis, NO_2 and PM_{10} are considered as the pollutants being investigated (due to the sparse observations for the other pollutants), whose data are obtained in two types, measured concentrations from the automatic networks and diffusion tubes, and modelled concentrations (DEFRA) at a 1 kilometre (km) resolution from an atmospheric dispersion model (AEA [1]). Disease data have been collected for $n = 1,207$ Intermediate Geographies (IG) in mainland Scotland, which have an average population of around 4,300 people. The disease data are from the Scottish neighbourhood statistics database (*http://www.sns.gov.uk/*), and comprise yearly numbers of admissions to non-psychiatric and non-obstetric hospitals in each IG with a primary diagnosis of respiratory disease (International Classification of Disease version 10 codes J00-J99). In addition to pollution and disease prevalence data, other covariate data such as socio-economic deprivation were also collected. Both the disease and covariate data are collected from 2007 to 2011, while the pollution data are from 2006 to 2010, to make sure that the exposure occurred before the hospital admissions. As the spatial patterns of disease and exposure are different in each year, here I consider the space-time correlation between disease and exposure rather than collapsing (averaging) across time and looking at space only.

A few key statistical challenges are covered in this thesis. The first one is how to improve the pollution predictions from a sparse network of monitoring stations. The second challenge lies in the predictions of more than one pollutants in a single model, which is required to allow the use of the correlation among pollutants to help improve the prediction of one pollutant by borrowing strength from the others. Another challenge is to consider the health effects of the exposure to multiple pollutants simultaneously, as the air people breathe is a mix of different pollutants. The last statistical challenge covered in this thesis is how to propagate exposure uncertainty into the investigation of

its health effects. This is important because the predicted exposures are always subject to uncertainty, such as the prediction error and the measurement error in measured data.

1.4 Thesis structure

I achieve the goal of this thesis by two main steps, improving the prediction of the spatially representative pollution concentrations and then investigating their impacts on health. More details are given as follows by introducing the structure of this thesis.

The remainder of this thesis is divided into six chapters. Chapter 2 provides an overview of the statistical methodology which is used in this thesis as well as the related literature. Chapter 3 is an initial impression about the air pollution health effects in mainland Scotland using a benchmark method. In chapter 4, I propose an integrated model for estimating the long-term health effects of NO_2, that fuses DEFRA and measured pollution data to provide improved predictions of areal level pollution concentrations and hence health effects. This is a single-pollutant health study. However, as the air we breathe contains a complex mixture of particle and gas phase pollutants, we are exposed to multiple pollutants simultaneously. These pollutants might act independently or in combination (in an additive, synergistic, antagonistic, or interactive manner) to affect human health. A traditional single pollutant health study fails to account for these combined effects of pollutant mixtures. Therefore, in chapter 5, I propose a multi-pollutant model which extends the single pollutant model in chapter 4, based on which the multi-pollutant concentrations can be predicted across mainland Scotland. The modelling carried out in chapter 5 will be used to provide pollution predictions for a study investigating the health effects of multi-pollutants in chapter 6. As the air pollution concentrations are spatially aggregated predictions from my pollution model, they are subject to variation. In addition, allowing the exposure uncertainty to be propagated into the investigation of its health impact is important in epidemiological studies. Therefore, in chapter 6, I also consider four approaches to adjust the exposure uncertainty. Finally, chapter 7 summarises the work contained within this thesis and discusses the implications for future research.

Chapter 2

Statistical background

This chapter introduces the statistical theory and methodologies used and developed, where section 2.1 introduces the normal linear model, in which the added variable plot is also introduced. Section 2.2 explores generalised linear models (GLMs) and their uses, with a particular focus on Poisson GLMs which are used in the spatial modelling approaches in this thesis. Spatial modelling and spatio-temporal modelling are introduced in Section 2.3, and these will form the basis of the methodology developed in chapters 4 and 5. Section 2.4 introduces Bayesian statistics, which is the statistical framework employed throughout this thesis, including the concepts of prior, posterior distributions, and methods of inference for Bayesian approaches.

2.1 The normal linear model

A linear regression model relates an observed quantity y to a number of other quantities, $z_1, z_2, ..., z_p$ as:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p + \epsilon. \tag{2.1}$$

$\beta_0, \beta_1, ..., \beta_p$ are parameters which are a key part of the systematic or structural part of the model and ϵ is an error term that accounts for uncertainties. y is the response variable, and $z_1, z_2, ..., z_p$ are explanatory variables.

The model parameters $\beta_j (j = 1, ..., p)$ describe how the mean value of y changes as the explanatory variables change, under the assumption that the underlying structure is linear. β_j can be interpreted as the amount of change in the mean value of y while z_j increases by one unit and the other explanatory variables are held fixed. The error term ϵ reflects the fact that data are subject to variation, from natural processes, measurement error and other sources.

If (2.1) is used for a set of n observations of the response and explanatory variables, the explicit form of the equations would be:

$$y_i = \beta_0 + \Sigma_{j=1}^{p} \beta_j z_{ij} + \epsilon_i, \quad i = 1, ..., n, \tag{2.2}$$

where for each i, y_i is the ith observation of the response, z_{ij} is the ith observation of the jth explanatory variable $(j = 1, 2, ..., p)$, and ϵ_i is the unobservable error corresponding to this observation. This set of n equations can be written in a compact form by,

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}, \tag{2.3}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & z_{11} & ... & z_{1p} \\ 1 & z_{21} & ... & z_{2p} \\ ... & ... & ... & ... \\ 1 & z_{n1} & ... & z_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{pmatrix}.$$

The description of Model (2.3) is completed by specifying a number of assumptions as follows, each of which needs to be considered and checked, where possible.

(A1) The relationship between the mean value of y and each z_j is linear if the other explanatory variables are held fixed.

(A2) The distribution of the error term is normal.

(A3) The variance of the error term is the same for all observations.

(A4) The error terms are independent.

Assumptions (A2) to (A4) can be written as $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, where σ^2 is a constant representing the variance of error term.

2.1.1 Model fitting

The least squares method is the oldest method used for estimation in the linear model. The error vector of the linear model (2.3) can be written as $(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})$, and the estimates of the model parameters $\boldsymbol{\beta}$ are obtained by minimising the sum of squared elements of this error vector. Such an estimator is called a least squares estimator (LSE) of $\boldsymbol{\beta}$. Formally, an LSE is

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}). \tag{2.4}$$

Provided $\boldsymbol{Z}^\top \boldsymbol{Z}$ has full column rank, the unique least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$.

The other two simple but important results which give the mean and variance of the least squares estimator are,

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{LS}] = \boldsymbol{\beta}, \mathrm{var}[\hat{\boldsymbol{\beta}}_{LS}] = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \sigma^2. \tag{2.5}$$

The first of these results tells us that the distribution of the least squares estimators is centred on the true value and it is unbiased. The second result expresses the precision of the estimates.

An unbiased estimate of σ^2 is given by,

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{y}}{n - p}, \tag{2.6}$$

where $\boldsymbol{I}$ is an $n \times n$ identity matrix and $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ is called the *hat-matrix*.

A measure of how well the model fits the data is called the *coefficient of determination*:

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}, \tag{2.7}$$

where $\bar{y}$ is the mean of n responses and $\hat{y}_i$ is the ith fitted value. Its range is $0 \leq R^2 \leq 1$, where values closer to 1 indicate a better fit.

2.1.2 Model checking

Before the linear model is used to draw conclusions, it is necessary to check that it does indeed fit the data well and provides a good description of the observed data. Notice that all four of the model assumptions (A1 to A4) can be expressed as statements about the error terms ϵ_i. Therefore, it is helpful to estimate these through the residuals $\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + ... + \hat{\beta}_p z_{ip})$. The following are some ways we can check the assumptions.

(1) If assumptions (A1), (A3) hold, a plot of the residuals against the fitted values $(\hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + ... + \hat{\beta}_p z_{ip})$ should show only random scatter, without any systematic patterns or change in the spread of the residuals.

(2) Assumption (A2) can be checked by drawing a histogram and a qq-plot. In the latter a straight line is expected.

(3) It is more difficult to check the assumption of independence as it depends on the type of data being modelled (e.g. time series, spatial, longitudinal). For example, it is useful to think carefully about the way in which the data have been collected to provide reassurance that there are no obvious ways in which dependence could arise.

If the model assumptions are not valid, one useful strategy is to seek transformations of the data (the response, the covariates or both) onto scales where the assumptions become valid. However, there are also cases where transformation cannot fully solve the problem. Sometimes this is because the type of data we are dealing with is intrinsically non-normal, for example in the form of counts. In that case it is better to use the

extension of linear models known as generalized linear models. This topic will be covered later in this chapter.

2.1.3 Added variable plot

In a linear model, for multiple regression, if one explanatory variable is highly correlated with other explanatory variables, the issue of collinearity will occur, which results in poor estimation of the influence of each explanatory variable separately.

If two explanatory variables (z_1, z_2) are perfectly linearly correlated ($cor(z_1, z_2) = \pm 1$), only z_1 or z_2 is needed to be in the linear model, because one of them contains all the information of the other with respect to the explanation of the variance of the response. Following this, an interesting question would be how to evaluate the necessity of adding an additional explanatory variable into the model.

The added variable plot is also referred to as a partial regression plot which is the most commonly used method for obtaining a graphical evaluation of the effect of adding an explanatory variable (say, X_i) to a model which already contains X_0,..., X_{i-1}. An added variable plot illustrates the incremental effect on the response of specific terms by removing the effects of all other terms. It is formed by: (1) Compute the residuals of regressing the response variable against the explanatory variables X_0,..., X_{i-1}; (2) Compute the residuals from regressing X_i against X_0,..., X_{i-1}; (3) Plot the residuals from (1) against the residuals from (2). If there is a pattern in this plot, then X_i should be added to the model (Ryan [110]).

As it is well known that the residuals from a linear model are uncorrelated to any explanatory variable in the model, I adopted this theory to deal with the multicollinearity in my study of multiple pollutants health effects. I regress one pollutant against another, e.g. regressing PM_{10} against NO_2. Then I use the residuals of this model, which are uncorrelated with NO_2 and represent the remaining signal of PM_{10} which cannot be explained by NO_2, and the NO_2 data in a single disease model without causing any multicollinearity issues, so as to investigate the health effects of exposure to both NO_2 and PM_{10} simultaneously.

2.2 Generalised linear models

The assumptions underlying a linear model are that (at least to a good approximation) the errors are normally distributed, the error variances are constant and independent of the mean, and the systematic effects combine additively (linearity). When data for a coninuous characteristic cover only a limited range of values, these assumptions may be justifiable. However, these assumptions could be far from being satisfied when the response is measured on a ratio scale or takes the form of a set of counts. Therefore, a generalised linear model (GLM) is a natural extension to the linear model, which allows the response variable, y, to be one from a set of independent random variables from any exponential family distribution, f. A random variable y belongs to the exponential family of distributions if its probability mass function (discrete) or probability density function (continuous) can be written in the canonical form:

$$p(y \mid \theta, \phi) = \exp[(y\theta - b(\theta))/a(\phi) + c(y, \phi)] \tag{2.8}$$

for some functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. Members of this exponential family include several important distributions: Gaussian, Binomial, Exponential and Poisson distributions.

A generalised linear model takes the form:

$$\begin{aligned} y_i &\sim f(\theta_i) \quad i = 1, ..., n, \\ g(\theta_i) &= \eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}, \end{aligned} \tag{2.9}$$

where $\boldsymbol{z}_i$ is a covariate vector, $\boldsymbol{\beta}$ is the unknown regression parameters, $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}$ is known as the linear predictor, and $g(\cdot)$ is a monotonic invertible link function which does not depend on $f(\cdot)$. The commonly used link functions $g(\cdot)$ include log, square root and logit transformations. The linear model (2.1) is a special case of the GLM, in which the link function is the identity function $g(\theta_i) = \theta_i$ and $f(y_i \mid \theta_i) = \mathrm{N}(\theta_i, \sigma^2)$.

The Poisson GLM is used throughout this thesis, as the disease data in my study are counts. These response data can only take a non-negative value, so the log is a suitable

and commonly used link function in the GLM. The basic Poisson GLM is specified as follows:

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

In this study, Poisson GLMs are fitted under the Bayesian framework which is introduced in section 2.4.

2.3 Spatial methods

Spatial analysis methods are used in a wide range of fields, such as spatial economics, image processing, epidemiology and environmental science. In spatial problems, spatial data are any form of statistical data which have geographical locations attached, and they are classified into three main types: point-referenced data, areal data and point pattern data. They come from different spatial processes, namely, geostatistical, areal and point processes. The key concepts about both point-referenced and areal data which are the two types of spatial data in my research are introduced in this section.

2.3.1 Geostatistical data

A geostatistical process is the stochastic process

$$X(\boldsymbol{s}) : \boldsymbol{s} \in D, \tag{2.11}$$

where $X(\boldsymbol{s})$ is the random variable representing the stochastic process at location $\boldsymbol{s}$, D is a fixed subset of the p-dimensional space $\mathbb{R}^p$. In my study, I focus on $p = 2$, so $D \subset \mathbb{R}^2$. The locations $\boldsymbol{s}$ at which data could occur varies continuously over D. However, in my study, data are observed at a finite number (n) of locations which are denoted by $\boldsymbol{x} = \{x(\boldsymbol{s}_1), ..., x(\boldsymbol{s}_n)\}$. The corresponding random variables are denoted by $\boldsymbol{X} = \{X(\boldsymbol{s}_1), ..., X(\boldsymbol{s}_n)\}$. An example of point-referenced data would be the concentrations

of air pollution in Scotland recorded at a number of monitoring stations. The overall concentration pattern for all of Scotland could then be estimated based on the data obtained at these fixed monitoring stations.

Geostatistics tries to answer questions about modelling, identification and separation of small and large scale variations, prediction (or kriging) at unobserved sites and reconstruction of $\boldsymbol{X}$ across the whole space. The key challenge when modelling spatial data compared with some types of non-spatial data is dependence (correlation). Typically, geostatistical data usually display positive correlation such that the nearer in space two observations are, the more similar their values are likely to be. This correlation is caused by the variable of interest being affected by other unmeasured processes which are themselves spatially correlated. In the following, more concepts about a geostatistical process used in this thesis are introduced.

Covariance

Covariance functions are used to quantify and model the correlation between observations. The covariance function of $X(\boldsymbol{s}) : \boldsymbol{s} \in D$ is defined as

$$\begin{aligned} C_X(\boldsymbol{s},\boldsymbol{t}) &= \mathrm{Cov}[X(\boldsymbol{s}), X(\boldsymbol{t})] \\ &= \mathbb{E}[(X(\boldsymbol{s}) - \mu_X(\boldsymbol{s}))(X(\boldsymbol{t}) - \mu_X(\boldsymbol{t}))], \end{aligned} \tag{2.12}$$

where $\mu_X(\boldsymbol{s})$ is the theoretical mean/expectation of the stochastic process $\{X(\boldsymbol{s})\}$ at location $\boldsymbol{s}$, and $\mu_X(\boldsymbol{t})$ at location $\boldsymbol{t}$. The covariance measures the strength of the linear dependence between $X(\boldsymbol{s})$ and $X(\boldsymbol{t})$.

The variance function of $\{X(\boldsymbol{s})\}$ is the special case of the covariance with $\boldsymbol{s} = \boldsymbol{t}$, which is

$$\begin{aligned} \mathrm{Var}[X(\boldsymbol{s})] &= C_X(\boldsymbol{s},\boldsymbol{s}) \\ &= \mathrm{Cov}[X(\boldsymbol{s}), X(\boldsymbol{s})] \\ &= \mathbb{E}[(X(\boldsymbol{s}) - \mu_X(\boldsymbol{s}))^2] \\ &= \sigma_X^2(\boldsymbol{s}), \end{aligned} \tag{2.13}$$

Semi-variogram

In geostatistics, the semi-variogram is used in exploratory data analysis to identify if there is any spatial correlation in the data. The semi-variogram of a geostatistical process $X(\boldsymbol{s}) : \boldsymbol{s} \in D$ is a function given as

$$\gamma_X(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{2}\text{Var}[X(\boldsymbol{s}) - X(\boldsymbol{t})], \tag{2.14}$$

which measures the variance of the difference in the process at two spatial locations $\boldsymbol{s}$ and $\boldsymbol{t}$.

The relationship between the semi-variogram and covariance is given as follows:

$$\begin{aligned} \gamma_X(\boldsymbol{s}, \boldsymbol{t}) &= \frac{1}{2}\text{Var}[X(\boldsymbol{s}) - X(\boldsymbol{t})] \\ &= \frac{1}{2}\text{Cov}[X(\boldsymbol{s}) - X(\boldsymbol{t}), X(\boldsymbol{s}) - X(\boldsymbol{t})] \\ &= \frac{1}{2}\{C_X(\boldsymbol{s}, \boldsymbol{s}) + C_X(\boldsymbol{t}, \boldsymbol{t}) - 2C_X(\boldsymbol{s}, \boldsymbol{t})\}, \end{aligned} \tag{2.15}$$

Let $\boldsymbol{t} = \boldsymbol{s}+\boldsymbol{h}$, then $\boldsymbol{h}$ is called the spatial lag or displacement. In addition, we assume that geostatistical process $X(\boldsymbol{s}) : \boldsymbol{s} \in D$ is weakly stationary (1st moment and autocovariance do not vary with respect to time), then,

$$\begin{aligned} \text{Cov}[X(\boldsymbol{s}), X(\boldsymbol{t})] &= \text{Cov}[X(\boldsymbol{s}), X(\boldsymbol{s} + \boldsymbol{h})] \\ &= C_X(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}) \\ &= C_X(\boldsymbol{h}). \end{aligned} \tag{2.16}$$

With this, the semi-variogram can be simplified to

$$\begin{aligned} \gamma_X(\boldsymbol{s},\boldsymbol{t}) = \gamma_X(\boldsymbol{h}) &= C_X(\mathbf{0}) - C_X(\boldsymbol{h}) \\ &= \sigma_X^2 - C_X(\boldsymbol{h}), \end{aligned} \tag{2.17}$$

and additionally assuming that the covariance between values of $X(\boldsymbol{s})$ at any two locations depends only on the distance between them (isotropy), we get,

$$\gamma_X(h) = \sigma_X^2 - C_X(h), \tag{2.18}$$

where $h = \|\boldsymbol{h}\|$ denotes the length of the lag vector $\boldsymbol{h}$ as measured by its Euclidean distance. In two dimensional geostatistics, $\boldsymbol{h} = (h_1, h_2)$, then we have $\|\boldsymbol{h}\| = \sqrt{(h_1^2 + h_2^2)}$. Therefore, the semi-variogram can be calculated given the covariance function.

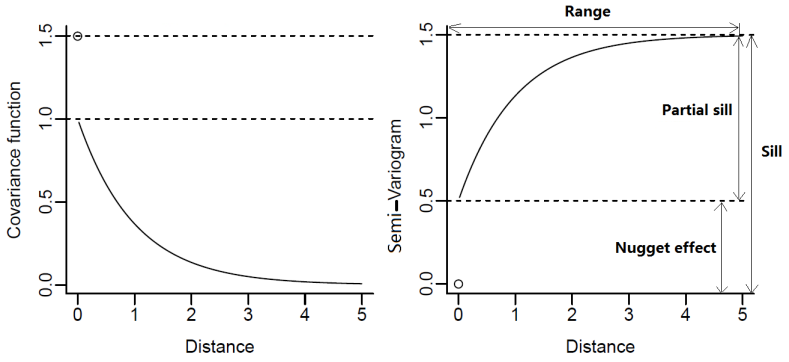
The most commonly used covariance functions include exponential, Gaussian, power exponential, spherical, wave and linear function (Diggle and Ribeiro [36]). A broad class of covariance models is the Matérn family functions, details of which can be seen in Matérn [86]. As exponential covariance has been commonly used in the existing literature (see e.g. Sahu et al. [112], Pannullo et al. [97] and Berrocal et al. [11]), in this study, it is also used in exploratory analysis to test any spatial correlation in pollution observations after accounting known covariates. I acknowledge the simplification associated with choosing the exponential covariance structure, however, other members of the Matérn family of covariance functions can be chosen. More details about exponentrial covariance function are presented as follows.

The exponential covariance function is

$$C_X(h) = \begin{cases} \sigma^2 \exp\left(-h/\phi\right), & h > 0; \\ \tau^2 + \sigma^2, & h = 0, \end{cases} \tag{2.19}$$

and the associated semi-variogram is

FIGURE 2.1: Exponential covariance function and associated semi-variogram.



$$\gamma_X(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-h/\phi)), & h > 0; \\ 0, & h = 0, \end{cases} \tag{2.20}$$

where the parameter $\tau^2 > 0$ is the nugget effect which is the limiting value of the semi-variogram as $h \to 0$. This parameter represents measurement error, or the spatial variability on a smaller scale than the distance between the two closest points in the sampling region (Diggle and Ribeiro [36]). On the other hand, while $h \to \infty$, the limiting value of the semi-variogram $(\tau^2 + \sigma^2)$ is called the sill. The parameter $\sigma^2 > 0$ is the partial sill which is also equal to the sill minus the nugget effect. The parameter ϕ is a range parameter measuring how quickly the covariance decays to zero. Another important parameter about the semi-variogram is the range, which is the distance beyond which locations will be spatially independent. Hence, responses at locations separated by distances greater than the range are spatially uncorrelated. For semi-variograms which reach their sill asymptotically, the effective, or practical range can be identified. One definition of the effective range is provided in Cressie [31] who proposed that the effective range is the distance at which the semi-variogram reaches 95% of its sill. An example of the exponential covariance function $C_X(h)$ and semi-variogram $\gamma_X(h)$ can be seen in Figure 2.1.

Spatial correlation investigation

The semi-variogram can be used to investigate the presence of spatial correlation in geostatistical data. Suppose we have a geostatistical process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ (e.g. the residuals from a linear regression), and observe the realization

$$\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^\top, \tag{2.21}$$

the semi-variogram for $Y(\boldsymbol{s})$ is defined by

$$\begin{aligned} \gamma_Y(\boldsymbol{s}, \boldsymbol{t}) &= \frac{1}{2}\mathrm{Var}[Y(\boldsymbol{s}) - Y(\boldsymbol{t})] \\ &= \frac{1}{2}\mathbb{E}[(Y(\boldsymbol{s}) - Y(\boldsymbol{t}))^2] - \frac{1}{2}\mathbb{E}[Y(\boldsymbol{s}) - Y(\boldsymbol{t})]^2. \end{aligned} \tag{2.22}$$

In practice, once any trend among the geostatistical data has been removed, $\mathbb{E}[Y(\boldsymbol{s}) - Y(\boldsymbol{t})] = 0$. Therefore, a coarse estimate of the semi-variogram for $Y(\boldsymbol{s})$ is given by

$$\gamma_Y(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{2}[y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j)]^2, \quad for \quad each \quad i \neq j. \tag{2.23}$$

Assuming $Y(\boldsymbol{s})$ is stationary and isotropic, a plot of these quantities versus $h_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is called a variogram cloud. However, this plot can be very noisy and it is hard to see spatial structure from it. Therefore, the empirical semi-variogram and the binned empirical semi-variogram are usually used in practice. When the geostatistical data are evenly allocated across space throughout the spatial domain, the empirical semi-variogram can be used instead of the semi-variogram to test the spatial correlation among data, which is defined by

$$\begin{aligned} \hat{\gamma}_Y(h) &= \frac{1}{2 \mid N(h) \mid}\Sigma_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(h)}[y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j)]^2, \\ N(h) &= \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \|\boldsymbol{s}_i - \boldsymbol{s}_j\| = h\}, \end{aligned} \tag{2.24}$$

where $N(h)$ denotes the set of pairs of spatial locations at a distance h apart, and $\mid N(h) \mid$ denotes the number of points in this set.

However, for unevenly spaced data throughout the spatial domain, the size of $N(h)$ may be one for a number of observable distances h. In this case, the true variogram cannot be well estimated by the empirical semi-variogram, and the binned empirical

semi-variogram is prefered. Suppose we partition the space of distances into K intervals (bins)

$$I_k = (h_{k-1}, h_k], \quad k = 1, ..., K, \tag{2.25}$$

where $0 = h_0 < h_1 < ... < h_K$. Let $h_k^m = (h_{k-1} + h_k)/2$ denote the midpoint of the interval, the pairs of distances in each interval is given by

$$N(h_k) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \in I_k\}, \tag{2.26}$$

and the binned empirical semi-variogram is

$$\hat{\gamma}_Y(h_k^m) \quad = \quad \frac{1}{2 \mid N(h_k) \mid} \Sigma_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(h_k)} [y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j)]^2, \tag{2.27}$$

where 30 pairs per bin is one rule of thumb to define I_k (Journel and Huijbregts [64]).

One way to assess the presence of spatial correlation is to plot the semi-variogram, and overlay on top the upper and lower limits for the set of semi-variograms that would have occurred under independence. These limits are computed using Monte Carlo methods and are often called a Monte Carlo envelope (Diggle and Ribeiro [36]). If the estimated semi-variogram from the data lie completely inside the envelope, then the data contain no substantial spatial correlation.

Geostatistical model

In my study, the geostatistical process, $\{X(\boldsymbol{s})\}$, is assumed to come from a stationary isotropic Gaussian process as the residuals of pollution observations are expected to have no spatial pattern after accounting for the covariates (e.g. monitoring site environment, temperature). It can be commonly modelled in an additive form, which decompose its variation into

$$X(\boldsymbol{s}) = \mu_X(\boldsymbol{s}) + \epsilon_X(\boldsymbol{s}), \tag{2.28}$$

where $\mu_X(\boldsymbol{s}) = \mathbb{E}[X(\boldsymbol{s})]$ is the spatially varying mean used to capture most of the spatial variation in the process $\{X(\boldsymbol{s})\}$. $\epsilon_X(\boldsymbol{s})$ is a spatial geostatistical error process used to capture the small scale correlation in $\{X(\boldsymbol{s})\}$. For the stationary Gaussian model, the parameters to be estimated are the mean $\mu_X(\boldsymbol{s})$ and any additional parameters which define the covariance structure of the data. Typically, these include the nugget effect, partial sill, and decay (or range) parameter.

A sensible method to complete the specification of the geostatistical model is to model the mean $\mu_X(\boldsymbol{s}) = \boldsymbol{z}(\boldsymbol{s})^\top \boldsymbol{\beta}$ as a linear combination of p covariates, that is $\boldsymbol{z}(\boldsymbol{s}) = (1, z_2(\boldsymbol{s}), ..., z_p(\boldsymbol{s}))$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. For n points, we can write down in compact matrix form,

$$\boldsymbol{\mu}_X = (\mu_X(\boldsymbol{s}_1), ..., \mu_X(\boldsymbol{s}_n)) = \boldsymbol{Z}\boldsymbol{\beta}, \tag{2.29}$$

where $\boldsymbol{Z}$ is the $n \times p$ design matrix of covariates for all n locations.

The error term of the model, $\boldsymbol{\epsilon}_X = (\epsilon_X(\boldsymbol{s}_1), ..., \epsilon_X(\boldsymbol{s}_n))$ is modelled as

$$\boldsymbol{\epsilon}_X = \mathrm{N}(\boldsymbol{0}, \Sigma(\boldsymbol{\theta})), \tag{2.30}$$

where in my study $\Sigma(\boldsymbol{\theta}) = \sigma^2 \exp(-\boldsymbol{D}/\lambda) + \tau^2 \boldsymbol{I}$ is the $n \times n$ covariance matrix given by an exponential correlation function of distance, in which $\boldsymbol{\theta}$ represents the covariance parameters $(\sigma^2, \lambda, \tau^2)$, and $\boldsymbol{D}$ is the $n \times n$ Euclidean distance matrix between the data locations, σ^2 represents the partial sill, τ^2 is the nugget effect and λ is the spatial range parameter. With these assumptions, the Gaussian geostatistical model can be written as,

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{Z\beta}, \Sigma(\boldsymbol{\theta})). \tag{2.31}$$

The parameters of Model (2.31) can be estimated by maximising the likelihood. An algorithm for maximisation of the log-likelihood proceeds as follows.

The log-likelihood function of Model (2.31) is given by,

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi) &= -\frac{1}{2}\{n \log(2\pi) + \log\{|\, \sigma^2 \exp(-\boldsymbol{D}/\lambda) + \tau^2 \boldsymbol{I}\,|\} \\ &+ (\boldsymbol{X} - \boldsymbol{Z\beta})^\top (\sigma^2 \exp(-\boldsymbol{D}/\lambda) + \tau^2 \boldsymbol{I})^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta})\} \end{aligned} \tag{2.32}$$

maximisation of which yields the maximum likelihood estimates of the model parameters. In order to make the estimation easier, we parametrise to $\nu^2 = \tau^2/\sigma^2$, where ν^2 is called the noise to signal ratio. I also denote $\boldsymbol{V} = \exp(-\boldsymbol{D}/\lambda) + \nu^2 \boldsymbol{I}$, so that $\Sigma(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{V}$. Given $\boldsymbol{V}$, the log-likelihood function is maximised at

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\boldsymbol{V}) &= (\boldsymbol{Z}^\top \boldsymbol{V}^{-1} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{V}^{-1} \boldsymbol{X} \\ \hat{\sigma}^2(\boldsymbol{\beta}, \boldsymbol{V}) &= n^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta})^\top \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta})^\top. \end{aligned} \tag{2.33}$$

In practice, as the maximum likelihood estimator for σ^2 is biased, so the alternative below is used,

$$\acute{\sigma}^2(\boldsymbol{\beta}, \boldsymbol{V}) = (n-p)^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta})^\top \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta})^\top \tag{2.34}$$

where p is the number of parameters in the mean model (2.29).

By substituting the estimates $\left(\hat{\boldsymbol{\beta}}(\boldsymbol{V}), \hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \boldsymbol{V})\right)$ into the log-likelihood function (2.32), we obtained the profile likelihood (or reduced likelihood) for (ϕ, ν^2):

$$L_0(\nu^2, \phi) = -\frac{1}{2}\{n \log(2\pi) + n \log \hat{\sigma}^2(\boldsymbol{V}) + \log |\boldsymbol{V}| + n\} \tag{2.35}$$

This is optimised numerically with respect to ϕ and ν^2, then the estimates $(\hat{\phi}, \hat{\nu}^2)$ are obtained, followed by back substitution to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$.

Spatio-temporal geostatistical model

When the geostatistical data contain more than one time period, a spatio-temporal geostatistical model is required. In this thesis, I focus on the spatio-temporal pollution model proposed by Sahu et al. [112], which is used in chapter 4 for a comparison to my proposed model. The model has the general form:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{O}_t + \boldsymbol{\epsilon}_t \quad t = 1, ..., T, \\ \boldsymbol{O}_t &= \rho \boldsymbol{O}_{t-1} + \boldsymbol{Z}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t \quad t = 2, ..., T, \end{aligned} \tag{2.36}$$

where $\boldsymbol{X}_t$ denotes the vector of measured pollution data in year t. These noisy data are modelled as a linear combination of the true values $\boldsymbol{O}_t$ and independent (white noise) errors $\boldsymbol{\epsilon}_t$. The true values are modelled with a first order autoregressive component $(\rho \boldsymbol{O}_{t-1})$, a regression component $(\boldsymbol{Z}_t \boldsymbol{\beta})$ where $\boldsymbol{Z}_t$ is the covariate matrix and $\boldsymbol{\beta}$ is a vector of regression parameters, and a spatial autocorrelation component $\boldsymbol{\eta}_t$. $\boldsymbol{\eta}_t$ is modelled independently for each time period, and is given a multivariate Gaussian prior with mean zero and an exponential correlation matrix.

2.3.2 Areal data

Denote a partition of n distinct regions as $\{\boldsymbol{B}_i : i = 1, ..., n\}$ and D is the region interest, such that $\bigcup_{i=1}^{n} \boldsymbol{B}_i = D$, $\boldsymbol{B}_i \bigcap \boldsymbol{B}_j = \emptyset$ for each $i \neq j$, then an areal process is the stochastic process

$$\{X(\boldsymbol{B}_i) : i = 1, ..., n\} \tag{2.37}$$

2.3.2.1 Spatial association in areal data

Similar to geostatistical data, measuring and then modelling the spatial association in areal data is also very important. The most common statistic to measure spatial correlation for areal data is Moran's I (Moran [91]) which is defined as,

$$I = \frac{n\Sigma_{i=1}\Sigma_{j=1}w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{w_{..}\Sigma_{i=1}^{n}(X_i - \bar{X})^2} \tag{2.38}$$

where $\boldsymbol{X} = (X_1, ..., X_n)$ denotes the areal process, w_{ij} is the element of a proximity matrix $\boldsymbol{W}$ and $w_{..} = \Sigma_i\Sigma_j w_{ij}$. The proximity matrix $\boldsymbol{W}$ defines how the n distinct regions $\{\boldsymbol{B}_i : i = 1, ..., n\}$ are potentially related to one another, the elements of which can be continuous (e.g. how far apart are the different regions from one another) or discrete (e.g. which regions are neighbours). In my study, I use the latter to define $\boldsymbol{W}$ and w_{ij} is equal to one if areas (i, j) share a common border, and is zero otherwise.

A permutation test can be used to test whether there is any spatial association in the areal data (H_0 - no spatial association; H_1 - some spatial association). To conduct this test, I compute the observed Moran's I test statistic, I_{obs} first, then calculate Moran's I statistics $(I_1, ..., I_k)$ based on K different random permutations of the areal dataset. Finally, the estimated two-sided p-value for the test is given as,

$$\frac{2}{K+1}\Sigma_{k=1}^{K}I(I_k > \mid I_{obs} \mid) \tag{2.39}$$

2.3.2.2 Spatial modelling for areal data

In my study, the disease data are counts of the numbers of disease cases occurring in each areal unit (IG), and thus Poisson log-linear models are typically used for the analysis. Denote the observed and expected numbers of respiratory hospital admissions in each IG by $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{E} = (E_1, \ldots, E_n)$, respectively, the latter of which is included in the regression model as an offset term. The expected counts are usually calculated by $E_k = \Sigma_{j=1}^{m}N_{jk}r_j$, where N_{jk} is the population in area k in strata j, r_j is the rate of

disease for strata j. The vector of p covariates for area k is denoted by $\mathbf{b}_k$, and x_k is the pollution concentration in area k. The spatial disease model is given by,

$$\begin{aligned} Y_k \mid E_k, R_k &\sim \text{Poisson}(E_k R_k) \qquad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \mathbf{b}_k^\top \boldsymbol{\alpha} + x_k \beta + \phi_k, \end{aligned} \tag{2.40}$$

where R_k denotes the overall risk of disease in area k, and a value of 1 corresponds to observing as many admissions as you expect from the population demographics (e.g. $E[Y_k] = E_k$). Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ are unknown covariate parameters, while β is an unknown parameter which quantifies the relationship between pollution and respiratory ill health. The last term in the model is a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ capturing the overdispersion and spatial correlation remaining in the disease data after adjusting for the covariates.

The random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are commonly modelled by the class of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model. Spatial correlation between the random effects is determined by a binary $n \times n$ neighbourhood matrix $\boldsymbol{W}$. Based on this neighbourhood matrix, the most common models for $\boldsymbol{\phi}$ include intrinsic autoregressive (IAR), convolution or BYM model, as well as those proposed by Cressie [31] and Leroux et al. [80].

Intrinsic CAR

The simplest CAR prior is the intrinsic autoregressive (IAR) model, which was proposed by Besag et al. [15] and has full conditional distribution $f(\phi_k \mid \boldsymbol{\phi_{-k}})$ given by

$$\phi_k \mid \boldsymbol{\phi}_{-k}, \boldsymbol{W}, \nu^2 \sim \text{N}\left(\frac{\Sigma_{j=1}^n w_{kj} \phi_j}{\Sigma_{j=1}^n w_{kj}}, \frac{\nu^2}{\Sigma_{j=1}^n w_{kj}} \right). \tag{2.41}$$

where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$. The conditional expectation of ϕ_k is represented by the mean of the random effects in neighbouring areas, and the conditional variance is inversely proportional to the number of neighbours. This variance structure indicates that the more neighbours an area has, the more information there is in the data

about the value of its random effect, which is obviously not the case if data contain weak to no spatial correlation. Therefore, this model is only appropriate for strong spatial correlation structures.

BYM

The convolution or BYM model, which was also proposed by Besag et al. [15], is given by

$$\begin{aligned} \phi_k &= \theta_k + \Psi_k, \\ \theta_k \mid \sigma^2 &\sim \mathrm{N}(0, \sigma^2), \\ \Psi_k \mid \boldsymbol{\Psi}_{-k}, \boldsymbol{W}, \nu^2 &\sim \mathrm{N}\left(\frac{\Sigma_{j=1}^n w_{kj} \Psi_j}{\Sigma_{j=1}^n w_{kj}}, \frac{\nu^2}{\Sigma_{j=1}^n w_{kj}} \right). \end{aligned} \tag{2.42}$$

The different strength of the spatial correlation can be achieved by varying the relative sizes of these two components $(\boldsymbol{\theta}, \boldsymbol{\Psi})$. However, each data point is represented by two random effects, and hence only their sum is identifiable.

Cressie CAR

This model was proposed by Cressie [31] and Stern and Cressie [121], and is given by

$$\phi_k \mid \boldsymbol{\phi}_{-k}, \boldsymbol{W}, \nu^2, \rho \sim \mathrm{N}\left(\frac{\rho \Sigma_{j=1}^n w_{kj} \phi_j}{\Sigma_{j=1}^n w_{kj}}, \frac{\nu^2}{\Sigma_{j=1}^n w_{kj}} \right). \tag{2.43}$$

When ρ equals to zero the random effects are independent, and there is no reason for the conditional variance of ϕ_k to be inversely proportional to the number of neighbours, as they provide no information about ϕ_k.

Leroux CAR

This model was proposed by Leroux et al. [80] which Lee [73] suggested to be preferred as it produces consistently good results across a range of spatial correlation strengths. This model has been further explored by Macnab [83]. Its univariate full conditional

distribution is given by

$$\phi_k \mid \boldsymbol{\phi}_{-k}, \boldsymbol{W}, \nu^2, \rho \sim \mathrm{N}\left(\frac{\rho\Sigma_{j=1}^{n} w_{kj}\phi_j}{\rho\Sigma_{j=1}^{n} w_{kj} + 1 - \rho}, \frac{\nu^2}{\rho\Sigma_{j=1}^{n} w_{kj} + 1 - \rho}\right). \tag{2.44}$$

The conditional expectation is a weighted average of the random effects in neighbouring areas, while the conditional variance has a more attractive form. When $\rho = 1$, the model reduces to the intrinsic model, $\rho = 0$ the conditional variance is a constant which means that there is no longer any information about ϕ_k in the neighbouring random effects.

2.3.2.3 Spatio-temporal modelling for areal data

The spatial modelling approaches introduced in last section can be used on areal data with only one time period. However in some cases, data are collected across T time points at each of the n areal units, and spatio-temporal modelling approaches are required. Denote (Y_{kt}, E_{kt}) as the observed and expected numbers of disease cases in areal unit k during time period t, the spatio-temporal disease model used in my study was developed by Rushworth et al. [109], and is given by:

$$\begin{aligned} Y_{kt} \mid E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\ \ln(R_{kt}) &= \boldsymbol{b}_{kt}^T\boldsymbol{\alpha} + x_{kt}\beta + \phi_{kt}, \\ \boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1} &\sim \mathrm{N}\left(\gamma\boldsymbol{\phi}_{t-1}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right), t \in 2, ..., T, \\ \boldsymbol{\phi}_1 &\sim \mathrm{N}\left(\mathbf{0}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right). \end{aligned} \tag{2.45}$$

The risk of disease in areal unit k and time period t is denoted by R_{kt}, and is modelled by three components on the log-scale. The first is a vector of covariates, $\boldsymbol{b}_{kt}$, and $\boldsymbol{\alpha}$ are the corresponding regression parameters. x_{kt} is the pollution concentration in areal unit k and time period t while β is an unknown parameter used to quantify the relationship between pollution and health.

ϕ_{kt} is a random effect included to allow for any spatio-temporal autocorrelation remaining in the disease counts after the covariate effects have been accounted for. This space-time specification of random effects is chosen instead of other ones with separate effects for space and time (e.g. $\phi_k + \phi_t$) because model (2.45) can be fitted by the R package CARBayesST directly. Here $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{nt})$ denotes the vector of random effects for time period t, and is modelled by a multivariate first order autoregressive process with temporal autocorrelation parameter γ and variance ν^2. Spatial autocorrelation is induced in the random effects by the precision matrix, which is given by $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$ and corresponds to the conditional autoregressive (CAR) prior proposed by Leroux et al. [80]. Further details about the specification of this model is given in Rushworth et al. [109]. Note that this space-time specification of random effects is chosen model 2.45

2.4 Bayesian modelling

In a statistical model, the observed data are usually believed to have come from a probability model with a set of unknown parameters which are then estimated from the data. A commonly used approach to estimate these unknown parameters is the likelihood approach. The parameters are chosen to be the best estimates which maximise the likelihood function. Under this framework, it is assumed that the unknown true values of the model parameters are fixed.

An alternative to the likelihood approach is the Bayesian approach, which comes from Thomas Bayes (1702-1761). He published a paper "An essay towards solving a problem in the doctrine of chances" which included a form of Bayes Theorem. The approach was also independently developed by Laplace approximately 50 years later.

2.4.1 Bayes' theorem

Bayes' theorem can be expressed for random variables. If random variables $\boldsymbol{\theta}$ (model parameters) and $\boldsymbol{Y}$ (data) have probability density function $f(\boldsymbol{\theta})$ and $f(\boldsymbol{Y})$ respectively, then,

$$f(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \frac{f(\boldsymbol{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{Y})}, \tag{2.46}$$

which shows the rule to update model parameters $\boldsymbol{\theta}$ by using data $\boldsymbol{Y}$. In Bayesian statistics, the parameters are treated as random and can therefore have probability distributions $f(\boldsymbol{\theta})$ assigned to them. The prior of the parameters is the belief about the parameters before observing any data, which then can be updated to get its posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{Y})$ in light of the observed data, $\boldsymbol{Y}$, via the data likelihood $f(\boldsymbol{Y} \mid \boldsymbol{\theta})$. $f(\boldsymbol{Y})$ is the marginal distribution of the observed data, which is independent of model parameters $\boldsymbol{\theta}$. Therefore, the posterior distributions can instead be expressed up to a constant of proportionality as

$$f(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}). \tag{2.47}$$

2.4.2 The first stage of a Bayesian model: probability model

Before using data to estimate model parameters, a probability distribution from which the data are collected needs to be identified. This probability model forms the first stage of the Bayesian model. In my study, the Poisson distribution (see e.g. model (2.40)) has been used as a disease model because the disease data are counts of the numbers of disease cases occurring in each areal unit (IG) and time period. The Normal distribution has been used in the air pollution models (see e.g. model (4.3) and (5.1)).

2.4.3 The second stage of the Bayesian model: prior

The prior in Bayesian statistics is the previous knowledge or belief about the parameters $\boldsymbol{\theta}$ before observing the data. The prior distribution could be based on information from previous studies on similar data sets or an estimate from an expert, or it could simply be used to represent a position of prior ignorance (non-informative prior).

In practice, we usually choose $f(\boldsymbol{\theta})$ to be within a standard family of distributions to make posterior computations tractable. In certain situations, we can choose the prior

distribution to be conjugate to the likelihood, in which case the prior and posterior distribution will be from the same family. For example, Table 2.1 show the commonly used conjugate priors in my study.

TABLE 2.1: Commonly used conjugate priors.

Likelihood	**Model parameter**	**Conjugate prior distribution**
Multivariate normal	Mean	Multivariate normal
Normal	Variance	Inverse Gamma
Multivariate normal	Covariance matrix	Inverse-Wishart

In some cases we may have little or no intuition about the value of the parameter in advance of observing the data. Then we need to represent our lack of prior knowledge by assigning a weakly informative prior which will have a negligible effect on the posterior. In such cases, the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{Y})$ is driven by the observed data rather than the prior. Three commonly used weakly informative priors are used in my study. For a real value parameter θ_i, I use a Gaussian distribution with a very large variance (e.g. $\theta_i \sim \mathrm{N}(0, 1000)$). For a parameter on a specific interval, a uniform distribution on the entire possible range of values (e.g. $\theta_i \sim \mathrm{Uniform}(0, 1)$) is used. For a variance parameter, a weakly informative Inverse-Gamma can be used (e.g. $\theta_i \sim \text{Inverse-Gamma}(a = 0.001, b = 0.001)$). In practice, the use of this non-informative prior distribution for variance parameter needs to be considered carefully as inferences can become very sensitive to a, b in the model for data sets in which low values of variance are possible (Gelman [45]).

2.4.4 Using the data to update the prior: posterior distribution

According to Bayes theorem, the posterior density function is proportional to the prior density times the likelihood function, $f(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$. In this section, I show how to derive the posterior distributions in Bayesian statistics by deriving the posterior distributions for the common priors and likelihoods in this thesis shown in Table 2.1.

- Model:

$$\begin{aligned} y_1, ..., y_n \mid \theta &\sim \text{i.i.d. } \mathrm{N}(\theta, \sigma^2), \quad \sigma \text{ known} \\ \theta &\sim \mathrm{N}(\mu_0, \tau_0^2), \quad \mu_0, \tau_0 \text{ fixed constants.} \end{aligned} \tag{2.48}$$

• Posterior:

$$\begin{aligned} f(\theta \mid \boldsymbol{Y}) &\propto f(\theta) f(\boldsymbol{Y} \mid \theta) \\ &= \exp\left\{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right\} \prod_{i=1}^{n} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{\Sigma_{i=1}^n (y_i-\theta)^2}{\sigma^2}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{\Sigma_{i=1}^n y_i^2 - 2\theta\Sigma_{i=1}^n y_i + n\theta^2}{\sigma^2}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{n\theta^2 - 2n\theta\bar{y}}{\sigma^2}\right]\right\} \\ &\sim \mathrm{N}(\mu_n, \tau_n^2), \end{aligned} \tag{2.49}$$

where

$$\begin{aligned} \frac{1}{\tau_n^2} &= \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}; \\ \mu_n &= \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}. \end{aligned} \tag{2.50}$$

• Model:

$$\begin{aligned} y_1, ..., y_n \mid \theta &\sim \text{i.i.d. } \mathrm{N}(\theta, \sigma^2), \quad \theta \text{ known} \\ \sigma^2 &\sim \text{Invese-Gamma}(a, b), \quad a, b \text{ fixed constants.} \end{aligned} \tag{2.51}$$

• Posterior:

$$\begin{aligned} f(\sigma^2 \mid \boldsymbol{Y}) &\propto f(\sigma^2) f(\boldsymbol{Y} \mid \sigma^2) \\ &\propto (\sigma^2)^{-(a+1)} \exp\left(-b/\sigma^2\right) \times (\sigma^2)^{-n/2} \exp\left\{-\frac{\Sigma_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right\} \\ &= (\sigma^2)^{-(a+n/2+1)} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{\Sigma_{i=1}^n (y_i-\theta)^2}{2}\right]\right\} \\ &\sim \text{Invese-Gamma}\left(a + n/2, b + \frac{\Sigma_{i=1}^n (y_i-\theta)^2}{2}\right). \end{aligned} \tag{2.52}$$

• Model:

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{V} &\sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{V}), \quad \boldsymbol{\theta} \text{ known} \\ \boldsymbol{V} &\sim \text{Invese-Wishart}(\nu, \boldsymbol{\Phi}_{n\times n}), \quad \nu, \boldsymbol{\Phi}_{n\times n} \text{ fixed constants.} \end{aligned} \tag{2.53}$$

• Posterior:

$$
\begin{aligned}
f(\boldsymbol{V} \mid \boldsymbol{Y}) &\propto f(\boldsymbol{V}) f(\boldsymbol{Y} \mid \boldsymbol{V}) \\
&\propto \mid \boldsymbol{V} \mid^{-(\nu+n+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}_{n\times n}\boldsymbol{V}^{-1})\right\} \mid \boldsymbol{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\theta})^{\top}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\theta})\right\} \\
&\propto \mid \boldsymbol{V} \mid^{-(\nu+n+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}_{n\times n}\boldsymbol{V}^{-1})\right\} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((\boldsymbol{Y}-\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{\theta})^{\top}\boldsymbol{V}^{-1}\right)\right\} \\
&\propto \mid \boldsymbol{V} \mid^{-(\nu+n+2)/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\boldsymbol{\Phi}_{n\times n}+(\boldsymbol{Y}-\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{\theta})^{\top}\right]\boldsymbol{V}^{-1}\right)\right\} \\
&\sim \text{Invese-Wishart}\left(\nu+1, \boldsymbol{\Phi}_{n\times n}+(\boldsymbol{Y}-\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{\theta})^{\top}\right).
\end{aligned} \tag{2.54}
$$

2.4.5 Inference

In Bayesian modelling, the deviance information criterion (DIC) (Spiegelhalter et al. [119]) is usually used as a Bayesian measure of model fit that is penalised for complexity. It can be written as a function of the log likelihood,
DIC $= 2\log L\{E(\boldsymbol{\phi} \mid \boldsymbol{y}) \mid \boldsymbol{y}\} - 4E_{\boldsymbol{\phi}|\boldsymbol{y}}\{\log L(\boldsymbol{\phi} \mid \boldsymbol{y})\}$, where $\boldsymbol{\phi}$ denotes model parameters and $\boldsymbol{y}$ denotes observed data. After selecting a Bayesian model, the inference is based on the posterior distributions of the unknown model parameters. Some of the posterior distributions are straightforward to compute. For example, distributions with a conjugate prior usually have a posterior distribution which follows a standard distributional form (see for example the posteriors discussed in section 2.4.4). In many cases, the calculation of the posterior distribution is much more complex which commonly requires a numerical simulation to draw a sample of model parameter values from an approximation of the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{Y})$, so as to estimate the distribution of the model parameters.

2.4.5.1 Markov chain Monte Carlo simulation

Markov chain Monte Carlo (McMC) simulation is the most common simulation method used to draw samples from the distributions of unknown model parameters when the likelihood is tractable. It operates by sequentially sampling parameter values from a Markov chain whose stationary distribution is exactly the desired joint posterior distribution of interest. The Bayesian McMC computing in my study is accomplished using one of two basic algorithms, the Gibbs sampling algorithm (Geman and Geman [46],

Gelfand and Smith [43]) and Metropolis-Hastings (M-H) algorithm (Metropolis et al. [88], Hastings [54]).

Gibbs Sampling algorithm

Suppose the parameter $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_p\}$. To implement the Gibbs sampling algorithm, we must assume that samples can be generated from each of the full (or complete) conditional distributions $\{f(\theta_i \mid \boldsymbol{\theta}_{j\neq i}, \boldsymbol{Y}), i = 1, ..., p\}$ in the model. These full conditional distributions are available in closed form. Given the current state of the Markov chain is $\boldsymbol{\theta}^{(0)} = \{\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}\}$, the algorithm proceeds as follows:

For $(t = 1, ..., T)$, repeat:

1. Draw $\theta_1^{(t)}$ from $f(\theta_1 \mid \theta_2^{(t-1)}, ..., \theta_p^{(t-1)}, \boldsymbol{Y})$;

2. Draw $\theta_2^{(t)}$ from $f(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)} ..., \theta_p^{(t-1)}, \boldsymbol{Y})$;

.

.

.

p. Draw $\theta_p^{(t)}$ from $f(\theta_p \mid \theta_1^{(t)}, ..., \theta_{p-1}^{(t)}, \boldsymbol{Y})$.

Components could also be updated in random order (random sweep).

Metropolis Hastings algorithm

When the prior $f(\boldsymbol{\theta})$ and the likelihood $f(\boldsymbol{Y} \mid \boldsymbol{\theta})$ are not a conjugate pair, the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$ is not in closed form, and the Metropolis Hastings algorithm can be used to draw samples from the joint posterior distribution. We begin the Metropolis Hastings algorithm by specifying a candidate (or proposal) density $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$ which is a valid density function for every possible value of the conditioning variable $\boldsymbol{\theta}^{(t-1)}$, and satisfies $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)$ (Metropolis algorithm, a special case of the Metropolis Hastings algorithm). Given a starting value $\boldsymbol{\theta}^{(0)}$ at iteration $t = 0$, the algorithm proceeds as follows:

For $(t = 1, ..., T)$, repeat:

1. Draw $\boldsymbol{\theta}^*$ from $q(\cdot \mid \boldsymbol{\theta}^{(t-1)})$;

2. Compute the ratio $r = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta}^*)f(\boldsymbol{\theta}^*)}{f(\boldsymbol{Y}|\boldsymbol{\theta}^{(t-1)})f(\boldsymbol{\theta}^{(t-1)})}$

3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$; If $r < 1$, $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^*, & \text{with probability } r \\ \boldsymbol{\theta}^{(t-1)}, & \text{with probability } 1-r \end{cases}$

For both algorithms, for t sufficiently large (say, bigger than t_0), $\{\boldsymbol{\theta}^{(t)}, t = t_0 + 1, ..., T\}$ is a (correlated) sample from the true posterior, from which any posterior quantities of interest may be estimated. For example, we can use a sample mean to estimate the posterior mean,

$$\hat{E}(\theta_i \mid \boldsymbol{Y}) = \frac{1}{T - t_0} \Sigma_{t=t_0+1}^{T} \theta_i^{(t)}. \tag{2.55}$$

The time from $t = 0$ to $t = t_0$ is commonly known as the burn-in period. An appropriate t_0 is chosen to guarantee the convergence of the chain of $\theta_i^{(t)}, t = t_0 + 1, ..., T$ which will be discussed later. In this study, I also use this posterior sample to build a 95% credible interval (easily obtained by taking the 2.5th and 97.5th percentiles of the simulated posterior draws as the lower and upper bounds respectively) of the parameter θ_i, which captures the uncertainty of estimation. Note that an McMC chain is strongly autocorrelated and produces clumpy samples that are unrepresentative, in the short run, of the true underlying posterior distribution. One way to decrease autocorrelation is to thin the sample, using only every nth step. Therefore, in practice a posterior sample used for inference is usually a thinning sample from a completed posterior sample.

Consider an MCMC chain that is strongly autocorrelated. Autocorrelation produces clumpy samples that are unrepresentative, in the short run, of the true underlying posterior distribution. Therefore, if possible, we would like to get rid of autocorrelation so that the MCMC sample provides a more precise estimate of the posterior sample. One way to decrease autocorrelation is to thin the sample, using only every nth step. If we keep 50,000 thinned steps with small autocorrelation, then we very probably have a more precise estimate of the posterior than 50,000 unthinned steps with high autocorrelation. But to get 50,000 kept steps in a thinned chain, we needed to generate n*50,000 steps. With such a long chain, the clumpy autocorrelation has probably all been averaged out anyway! In fact, Link and Eaton show that the longer (unthinned) chain usually yields

better estimates of the true posterior than the shorter thinned chain, even for percentiles in the tail of the distribution, at least for the particular cases they consider.

Convergence monitoring and diagnosis

When an McMC algorithm has converged at time T, its output can be treated as coming from the true stationary distribution of the Markov chain for all $t > T$. Some researchers have attempted to establish conditions for convergence of various McMC algorithms under a rigorous mathematical framework (Roberts and Smith [107], Roberts and Tweedie [108], Meyn and Tweedie [89]). In this thesis, the convergence of McMC is checked by doing trace plots (should be no patterns) or by the Heidelberg and Welch Diagnostic (Heidelberger and Welch [57], Heidelberger and Welch [58]), which consists of two parts: a stationary portion test and a half-width test. The stationarity test assesses the stationarity of a Markov chain by testing the hypothesis that the chain comes from a covariance stationary process. The half-width test checks whether the Markov chain sample size is adequate to estimate the mean values accurately. Heidelberger and Welch [58] combined the method of Schruben [115] and Schruben et al. [114] to propose a comprehensive procedure for generating a confidence interval of prespecified width for the mean of a parameter when the chain has an *initial transient* (a state when the algorithm has not reached stationarity yet). The diagnostic is appropriate for the analysis of individual chains, and the procedure is given as follows.

Given an McMC chain $x_j : j = 1, ..., n$, the stationarity test of Schruben [115] and Schruben et al. [114] is applied to the chain, in which the null hypothesis of convergence is based on Brownian bridge theory and uses the Cramer-von-Mises test statistic (Cowles and Carlin [30])

$$\int_0^1 \left(\frac{T_{\lfloor nt \rfloor} - \lfloor nt \rfloor \bar{x}}{\sqrt{nS(0)}} \right)^2 dt \tag{2.56}$$

where

$$T_k = \begin{cases} 0, & k = 0 \\ \Sigma_{j=1}^{k} x_j, & k \geq 1 \end{cases} \tag{2.57}$$

and $S(0)$ is the spectral density evaluated at frequency zero. The spectral density is estimated from the second half of the original McMC chain. If the null hypothesis is rejected, then the first $0.1n$ of the samples are discarded and the test reapplied to the remaining chain. This processes is repeated until the test is either non-significant or 50% of the samples have been discarded, at which point the chain is declared to be non-stationary and the McMC chain needs to run longer. If convergence is not rejected in the final step, a half-width test is performed by computing the mean and associated $(1 - \alpha)100\%$ confidence interval. This test is passed if the half-width of the confidence interval is less than a user-specified level of accuracy ϵ, otherwise the test is failed (Cowles and Carlin [30]).

Chapter 3

The impact of air pollution on health

3.1 Introduction

The health impact of air pollution exposure has been widely recognised since the 1950's, as a result of the London smog in December 1952, which is estimated to have resulted in more than 3,000 excess deaths compared with previous years (Bell and Davis [9]). Recently, the adverse effects of air pollution on health have been widely investigated all over the world. In the USA, an expert elicitation by the US Environmental Protection Agency [133] reports an increase of 1% (range 0.4% to 1.8%) in annual all-cause deaths for a $1\mu gm^{-3}$ increase in the annual average of $PM_{2.5}$ exposure in the United States. In China, focusing on 17 Chinese cities, Chen et al. [27] found that an increase of $10\mu gm^{-3}$ in the two-day moving average SO_2 concentration was associated with 0.75% (95% posterior interval (PI), 0.47% to 1.02%), 0.83% (95% PI, 0.47% to 1.19%) and 1.25% (95% PI, 0.78% to 1.73%) increase of total, cardiovascular and respiratory mortality, respectively. In Europe the Dutch cohort study by Hoek et al. [59] found a 17% (95% confidence interval (CI), 24% to 78%) adjusted excess risk for all-cause mortality with a $10\mu gm^{-3}$ increase of background concentrations of black smoke (BS). In the UK, Elliott et al. [41] found significant associations between BS and SO_2 concentrations and mortality. Lee [74] reported that a $1.7\mu gm^{-3}$ increase in PM_{10} concentrations was associated with 6.6% additional hospital admissions across Scotland, while Lee et al. [76] showed that

long-term exposure (over 3 years) to PM_{10} and NO_2 was significantly associated with respiratory hospital admissions in Edinburgh and Glasgow while the risks for Aberdeen and Dundee were generally positive but nonsignificant.

The air pollutants commonly associated with ill health include $PM_{2.5}$, PM_{10}, NO_2, SO_2 and O_3, the effects of which on specific diseases can vary. For example, respiratory disease is often associated with the concentrations of PM_{10} (Lee [74], Lee et al. [76]), $PM_{2.5}$ (Tecer et al. [125]), NO_2 (Lee et al. [76]) and SO_2 (Elliott et al. [41]), while skin cancer has been associated with O_3 (Thormod et al. [128], Diepgen and Mahler [35]).

The disease and pollutants considered in this study are respiratory disease and NO_2 and PM_{10}, respectively. The disease data are counts of the numbers of respiratory hospital admissions within non-overlapping areal units. This is called a small-area ecological study design. The pollution data are from two sources: (a) measured data which are sparsely distributed in space; (b) modelled grid data from dispersion models. Therefore, the pollutant data and the disease data relate to different spatial scales, and the first thing we need to do before investigating the relationship between them is to convert the pollution data to a comparable small-area scale on which the disease data are aggregated. The simplest method for doing this is to compute the spatial mean concentrations over the modelled grid data lying within each small area, which ignores the monitoring data. A more complex approach is to firstly fuse measured data and modelled grid data before converting them into the small-area scale just as the first method does, which is expected to provide improved predictions of areal level pollution concentrations. The first method can be treated as a benchmark method to deal with the issue, as it is adopted in most of the existing research (see e.g. Maheswaran et al. [84], Lee [73], Lee [74], Warren et al. [136], Rushworth et al. [109]), and is therefore adopted in this chapter to get an initial impression of the impact of air pollution on health in Scotland.

Note that various aggregation functions for transferring spatial data into a single metric have been discussed by researchers (see e.g. Bruno and Cocchi [20]), however, the existing literature in the context of investigating air pollution health effects uses the average metric (mean or median) almost exclusively (e.g. Maheswaran et al. [84], Lee et al. [76], Lee [73], Rushworth et al. [109]). In this study, I investigate both spatial mean and maximum metrics as it may be that peak concentrations are more representative for the exposure. For example, NO_2 concentrations are usually higher near main roads, where

most of the exhaust fumes are produced, and peak concentrations are more suitable for the estimation of the real exposure if the population are dense next to main roads.

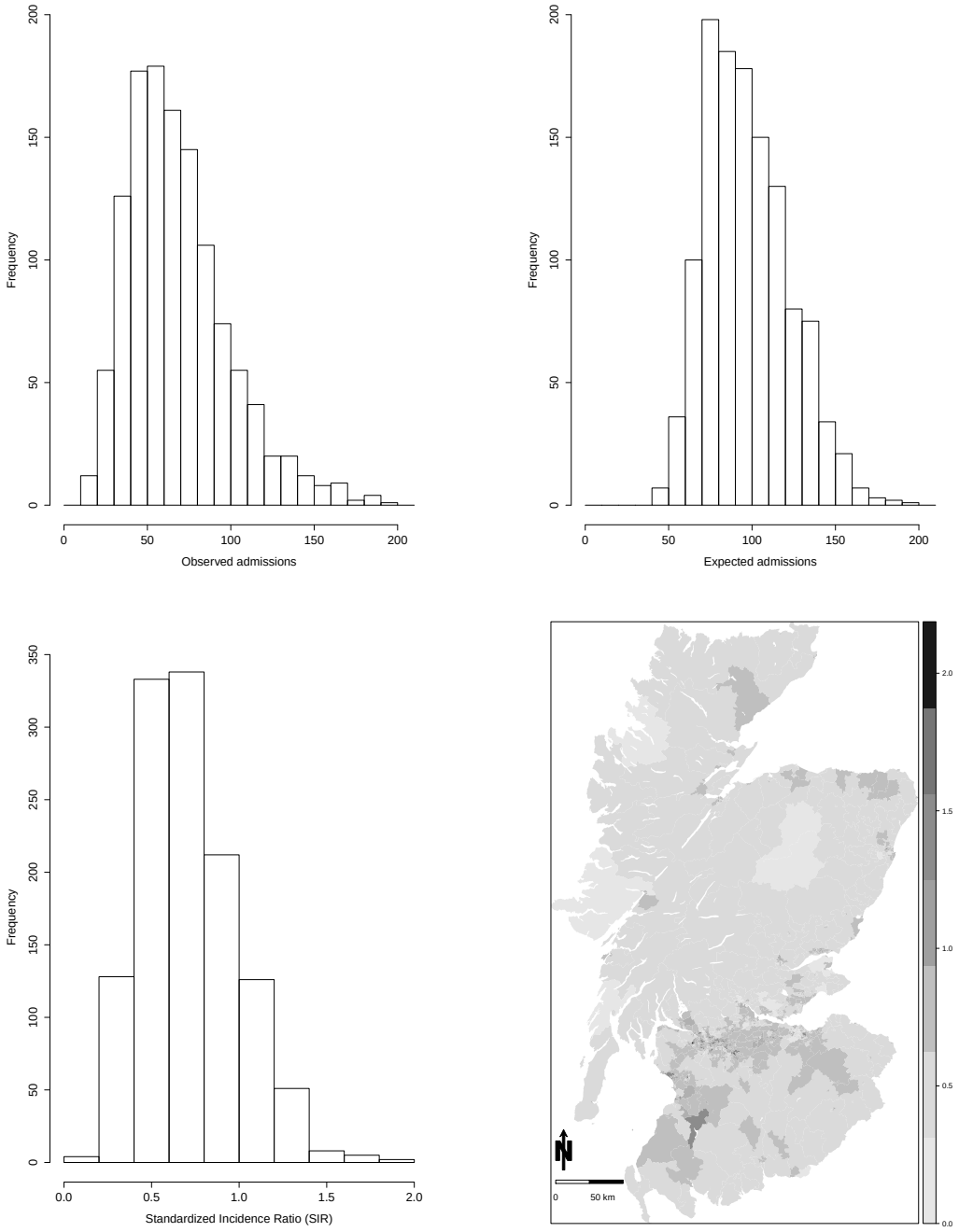
3.2 Data description

The data in this study region relate to the set of 1,207 Intermediate Geographies (IG) that comprise mainland Scotland, which each has an average population of around 4,300 people. The disease data analysed in this study are from the Scottish neighbourhood statistics database, whose website is *http://www.sns.gov.uk/*. The response variable is the numbers of admissions to non-psychiatric and non-obstetric hospitals in each IG in 2011 with a primary diagnosis of respiratory disease. These data are denoted by $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where Y_k denotes the count for area k. The number of admissions in an IG depends on its population size and demographic structure. Therefore I use age and sex as external variables to calculate the expected number of admissions in each IG based on standard hospital admission rates stratified by age (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+ years) and sex for the whole of Scotland. These rates can be obtained from the Information Services Division, which is part of the National Health Service in Scotland. The equation $E_k = \Sigma_{j=1}^{m} N_{jk} r_j$ indicates how to compute the expected counts, where E_k denotes the expected count in area k, N_{jk} is the population in area k in strata j, r_j is the rate of disease for strata j in Scotland.

The distribution of both observed and expected admissions in 2011 are shown in Figure 3.1, where the bottom panels show the histogram and spatial distribution of standardized incidence ratio (SIR) where $SIR_k = Y_k/E_k$. Figure 3.1 shows that overall the expected hospital admissions are over-estimated comparing to real observed hospital admissions, since the median of the former (about 100) is much higher than that from the latter (about 70). The spatial map of SIR shows that a higher standardized incidence ratio is found in Glasgow and Edinburgh cities (the set of small densely populated IGs in the lower middle part of the country) and those around Loch Doon (the darker area at bottom left).

The pollutants considered in this study are both NO_2 and PM_{10}. Strong relationships between respiratory diseases and NO_2 and PM_{10} have been demonstrated in related research, such as Oftedal et al. [95], Belanger et al. [8], Kattan et al. [66], Wiwanitkit

FIGURE 3.1: The distributions of the observed and expected admissions in 2011 and their standardized incidence ratio: top left is the observed hospital admissions, top right is the expected hospital admissions, two figures in the bottom are the corresponding SIR.



[140], Thishan Dharshana and Coowanitwong [126], Lee et al. [76] and Lee [74]. The modelled pollution data gridded to 1 km spatial resolution in this study are freely available, and can be downloaded from the Department for Environment Food and Rural Affairs (DEFRA) database (http://uk-air.defra.gov.uk/). I use pollution data for 2010 in this study rather than 2011, to make sure that the air pollution exposure occurred before the respiratory disease hospital admissions. However, these modelled grid data are only available for each 1 km grid square across the UK, which does not match the resolution of Intermediate Geographies. Therefore, I converted them to the Intermediate Geography scale by computing the spatial mean or maximum concentrations over the grid squares lying within each IG. The spatial distributions of NO_2 and PM_{10} in Scotland are shown in Figure 3.2. It is obvious that both NO_2 and PM_{10} concentrations are higher in the east of Scotland as well as in Glasgow and Edinburgh. The maps using spatial mean metric are much smoother than those from using spatial maximum metric.

A number of covariates corresponding to 2011 are considered to describe the spatial pattern in disease risk, including the measures of socio-economic deprivation: (a) the percentage of people living in each IG who are in receipt of Job Seekers Allowance (JSA), and (b) the natural log of median property price in an area (Logprice).

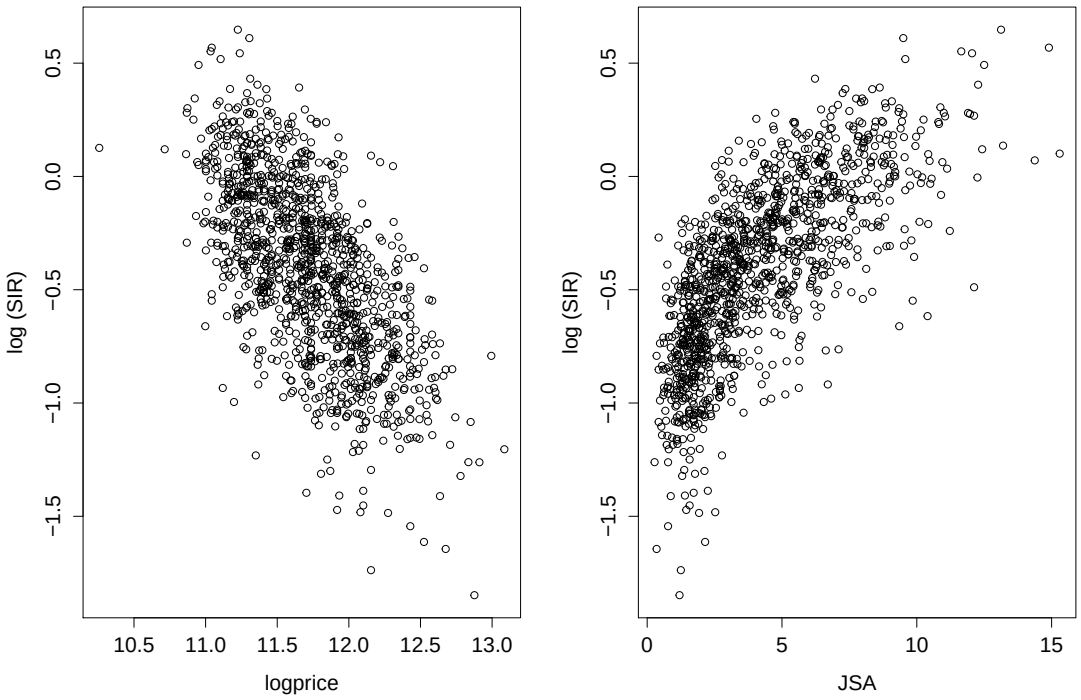
3.3 Exploratory analysis

The relationships between the covariate variables and respiratory hospital admissions are shown in Figure 3.3 where logSIR denotes the natural log of the SIR (the scale on which the covariates will be modelled). Figure 3.3 suggests a linear relationship between log(SIR) and the covariates Logprice. Note that the relationship between log(SIR) and JSA is only roughly linear, however, the linear relationship is assumed in my study so that it is easier to interpret.

Previous studies such as Lee et al. [76], Lee [73], Lee [74], have shown that disease data in contiguous areal units usually contain spatial correlations even after incorporating covariates, which is due to those confounding covariates that have not been identified. Therefore, I assessed the spatial correlation of the residuals from a non-spatial Poisson model in which the observed admissions counts are regressed against NO_2 or PM_{10}, the percentage of people living in each IG who are in receipt of Job Seekers Allowance, the

FIGURE 3.2: The spatial distributions of NO_2 and PM_{10} for 2010 in Scotland (unit: μgm^{-3}). Top left is based on using the mean gridded NO_2 concentrations in each IG, top right is based on using the max gridded NO_2 concentrations in each IG, bottom left is based on using the mean gridded PM_{10} concentrations in each IG, bottom right is based on using the max gridded PM_{10} concentrations in each IG.

FIGURE 3.3: Scatterplots of log respiratory disease SIR (log(SIR)) against Job Seekers Allowance (JSA) and log of median property price (logprice).



natural log of median property price in each IG, and the expected number of admissions as an offset term. Moran's I is adopted to test the spatial correlation of these model residuals. For both of the models with NO_2 or PM_{10}, Monte-Carlo simulation of Moran's I p-value equals 9.999e-05 indicating that the residuals of the non-spatial model contain a strong spatial correlation structure.

3.4 Methods

As it was shown in the previous section that the residuals of the non-spatial model contain a strong spatial correlation structure after accounting for the covariate effects, I use spatial models to model the disease data. These models are poisson log-linear models combined with conditional autoregressive (CAR) models to deal with the spatial correlation coming from this small-area ecological study design. The spatial pattern in the disease data are modelled by known covariates and random effects, the latter accounting for residual spatial correlation.

I denote the observed and expected numbers of respiratory hospital admissions in 2011 in each IG by $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{E} = (E_1, \ldots, E_n)$, respectively, the latter of which is included in the regression model as an offset term. The p covariates are denoted by $\boldsymbol{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$, where $\mathbf{b}_k$ is the vector of observations for area k. I denote the vector of NO_2 (or PM_{10}) concentrations in 2010 as $\mathbf{x} = (x_1, \ldots, x_n)$. In this chapter, I consider two model formulations, Gaussian and Poisson, to assess the robustness of my conclusions. The Gaussian regression model used in my study is given by eqn (3.1) while the Poisson regression model is given by eqn (3.2).

$$\ln\left(\frac{Y_k}{E_k}\right) \sim \mathrm{N}\left(\mathbf{b}_k^\top \boldsymbol{\alpha} + x_k\lambda + \phi_k, \sigma^2\right) \qquad \text{for } k = 1, \ldots, n, \tag{3.1}$$

$$\begin{aligned} Y_k \mid E_k, R_k &\sim \text{Poisson}(E_k R_k) \qquad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \mathbf{b}_k^\top \boldsymbol{\alpha} + x_k\lambda + \phi_k, \end{aligned} \tag{3.2}$$

where R_k denotes the overall risk of disease in area k, and a value of 1 corresponds to observing as many admissions as you expect given the population demographics (e.g. $E[Y_k] = E_k$). Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ are unknown covariate parameters, while λ is an unknown parameter which quantifies the relationship between NO_2 (or PM_{10}) and respiratory ill health. The last term in the model is a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ capturing the overdispersion and spatial correlation remaining in the disease data after adjusting for the covariates.

In disease mapping studies, the random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are commonly modelled by the class of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model. Spatial correlation between the random effects is determined by a binary $n \times n$ neighbourhood matrix W, whose jk^{th} element w_{jk} is equal to 1 if areas (i, j) are defined to be neighbours, and is 0 otherwise. The three most common ways to define areas (i, j) to be neighbours are: (i) they share a common border, (ii) their central points are within a fixed distance, (iii) one area is one of the h closest areas to another area in terms of distance. In this thesis, I utilize the common border specification which is the most standard. Four commonly used conditional autoregressive models, intrinsic autoregressive (IAR), convolution or BYM model, as well as those

proposed by Cressie [31] and Leroux et al. [80] are used in this chapter. They have been introduced in detail in chapter 2.

3.5 Results

For all the results presented in this section, inference is achieved using McMC simulation, where the Markov chain was burnt in for 20,000 iterations, after which convergence was assessed to have been reached, and then the remaining 30,000 iterations were used for the final results.

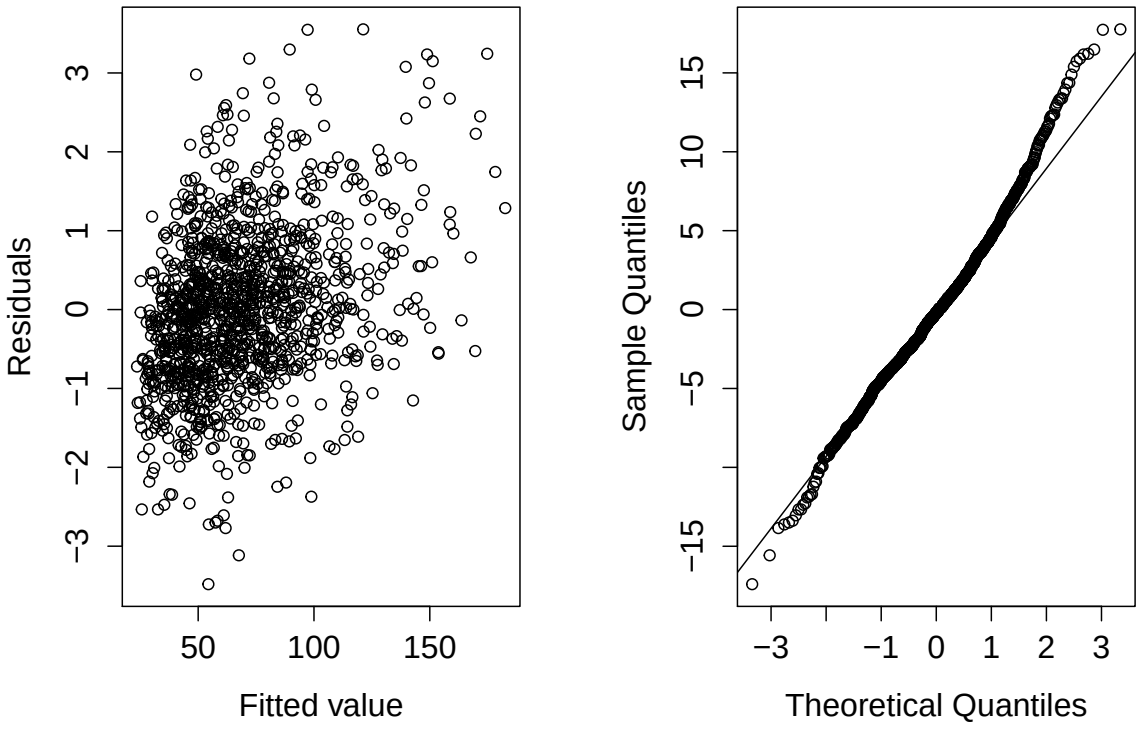
3.5.1 The effect of NO_2 on health

To assess the robustness of the conclusions, I estimate the association between air pollution and health (λ) using 8 models, which combine the likelihood models (3.1) and (3.2) with the prior models (2.41) to (2.44). The models are implemented in R using the package CARBayes proposed by Lee [75] which is freely available from the Comprehensive R Archive Network (CRAN,http://CRAN.R-project.org/package=CARBayes). CARBayes can fit the general exponential family Bayesian hierarchical model where the response data can be Binomial, Gaussian or Poisson.

The fit of the models was assessed by checking their residuals. Take an example from modelling maximum NO_2 using the Poisson model with the Leroux CAR prior. The left panel in Figure 3.4 shows the standardized residuals against the fitted values, in which no pattern is found. The right panel is a normal qq plot of the residuals to assess their normality, which shows that most of the points follow a linear trend suggesting the residuals of the model can be treated as normally distributed. Therefore, the model appears to be appropriate for the health data. Note that the Poisson distribution can be considered approximately Normal when its mean is larger than 20 (Central Limit theorem), and the responses in my data set are overall much higher than 20.

The output of the Poisson model with the Leroux CAR prior is shown in Table 3.1 (using mean NO_2 as pollution concentrations) and Table 3.2 (using maximum NO_2 as pollution concentrations), in which, according to the 95% credible intervals for model coefficients, there is evidence that the percentage of people living in each IG who are in

FIGURE 3.4: Model residuals from fitting a Poisson model for maximum NO_2 with the Leroux CAR prior.



receipt of Job Seekers Allowance (JSA) and the log median property price in an area all have a significant association with the number of hospital admissions due to respiratory disease, because their 95% credible intervals do not contain the neutral value, 0. Note that the regression coefficient for mean NO_2 in Table 3.1, 0.0031, represents the log increase of SIR for 1 μgm^{-3} increase of NO_2, and it is the same for Table 3.2.

Table 3.1 also shows that the mean NO_2 concentration in each IG is not significantly associated with respiratory disease hospital admissions due to respiratory diseases, as the corresponding 95% credible interval for λ contains 0. However, there is a significant association for maximum NO_2 as shown in Table 3.2.

In order to test the robustness of my conclusions, the remaining seven models were applied and their results are compared in Table 3.3 (mean NO_2) and Table 3.4 (max NO_2). These tables present a comparison of the estimated relative risk ($\exp(\lambda * sd)$) based on a standard deviation increase of NO_2 which is 6.84μgm^{-3} in my study.

TABLE 3.1: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for mean NO_2 with a Leroux random effect.

Variable	**Mean**	**2.5%**	**97.5%**
(Intercept)	1.4569	0.8440	2.0876
Mean NO_2	0.0030	-0.0007	0.0069
Logprice	-0.1873	-0.2382	-0.1366
JSA	0.0713	0.0636	0.0793
τ^2	0.1045	0.0913	0.1190
ρ	0.7900	0.6600	0.9000

TABLE 3.2: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for max NO_2 with a Leroux random effect.

Variable	**Mean**	**2.5%**	**97.5%**
(Intercept)	1.5984	1.0105	2.3006
Max NO_2	0.0050	0.0019	0.0078
Logprice	-0.2018	-0.2591	-0.1530
JSA	0.0707	0.0626	0.0782
τ^2	0.1033	0.0901	0.1180
ρ	0.7800	0.6500	0.8900

TABLE 3.3: Relative risk for a $6.84\mu gm^{-3}$ increase of NO_2 for eight models based on mean NO_2 data in each IG.

Models	**Relative risk**	**95%credible interval**	**DIC**
G.bymCAR	1.009	(0.983, 1.035)	-2250.267
G.iarCAR	1.009	(0.982, 1.036)	-228.072
G.lerouxCAR	1.016	(0.990, 1.043)	-254.780
G.cressieCAR	1.016	(0.990, 1.043)	-245.202
P.bymCAR	1.011	(0.984, 1.039)	9227.990
P.iarCAR	1.000	(0.973, 1.026)	9263.629
P.lerouxCAR	1.021	(0.995, 1.048)	9254.250
P.cressieCAR	1.017	(0.989, 1.043)	9261.201

According to Table 3.3, the relative risk of mean NO_2 in all models is not significant, as all the 95% credible intervals for the estimates contain the null risk of 1. This indicates that mean NO_2 does not have a significant influence on respiratory disease. This consistency in the estimation of relative risk also indicates the robustness of the relative risk.

Similarly, while the peak concentration in each IG is used to represent the exposure, the relative risk estimates across the eight models are robust (see Table 3.4), as the estimated relative risks are all significant and similar ranging from 1.021 to 1.034. All models indicate a significant association between the peak NO_2 concentrations in each

TABLE 3.4: Relative risk for a 6.84μgm^{-3} increase of NO_2 for eight models based on maximum NO_2 data in each IG.

Models	**Relative risk**	**95%credible interval**	**DIC**
G.bymCAR	1.023	(1.002, 1.046)	-729.099
G.iarCAR	1.023	(1.001, 1.045)	-228.573
G.lerouxCAR	1.028	(1.006, 1.050)	-260.555
G.cressieCAR	1.028	(1.005, 1.050)	-250.423
P.bymCAR	1.021	(1.001, 1.042)	9230.584
P.iarCAR	1.024	(1.002, 1.043)	9264.303
P.lerouxCAR	1.035	(1.013, 1.055)	9251.032
P.cressieCAR	1.031	(1.008, 1.052)	9258.653

IG and the respiratory disease hospital admissions. The relative risk in these models can be interpreted as: with a 6.84μgm^{-3} increase of peak NO_2 concentration, hospital admissions related to respiratory disease in each IG will increase by about 2.6% (ranging from 2.1% to 3.5%).

3.5.2 The effect of PM_{10} on health

While investigating the effect of PM_{10}, the results of a Poisson model with a LerouxCAR prior (an example which is consistent with NO_2) are shown in Table 3.5 and Table 3.6. Similarly, the normality test of model residuals for maximum PM_{10} is shown in Figure 3.5, in which both the standardized residuals against the fitted values plot and the normal qq plot of the residuals indicate the model is appropriate for the health data.

Table 3.5 and Table 3.6 show that the percentage of people living in each IG who are in receipt of Job Seekers Allowance (JSA) and the log median property price in an area all are associated with the number of hospital admissions due to respiratory disease. Both mean and maximum PM_{10} concentrations in each IG are associated with respiratory disease hospital admissions, as the corresponding 95% credible intervals of λ don't contains 0. Note that the regression coefficient for mean PM_{10} in Table 3.5, 0.0281, represents the log increase of SIR for 1 μgm^{-3} increase of PM_{10}, and it is the same for Table 3.6.

I also assess the robustness of the relative risk of PM_{10} using 8 models, which combine the likelihood models (3.1) and (3.2) with the prior models (2.41) to (2.44). The estimated relative risk from these 8 models, based on a standard deviation increase of PM_{10}, which is 1.872μgm^{-3} in my study, are shown in Table 3.7 and Table 3.8. According to Table 3.7,

FIGURE 3.5: Model residuals from fitting a Poisson model for maximum PM_{10} with the Leroux CAR prior.

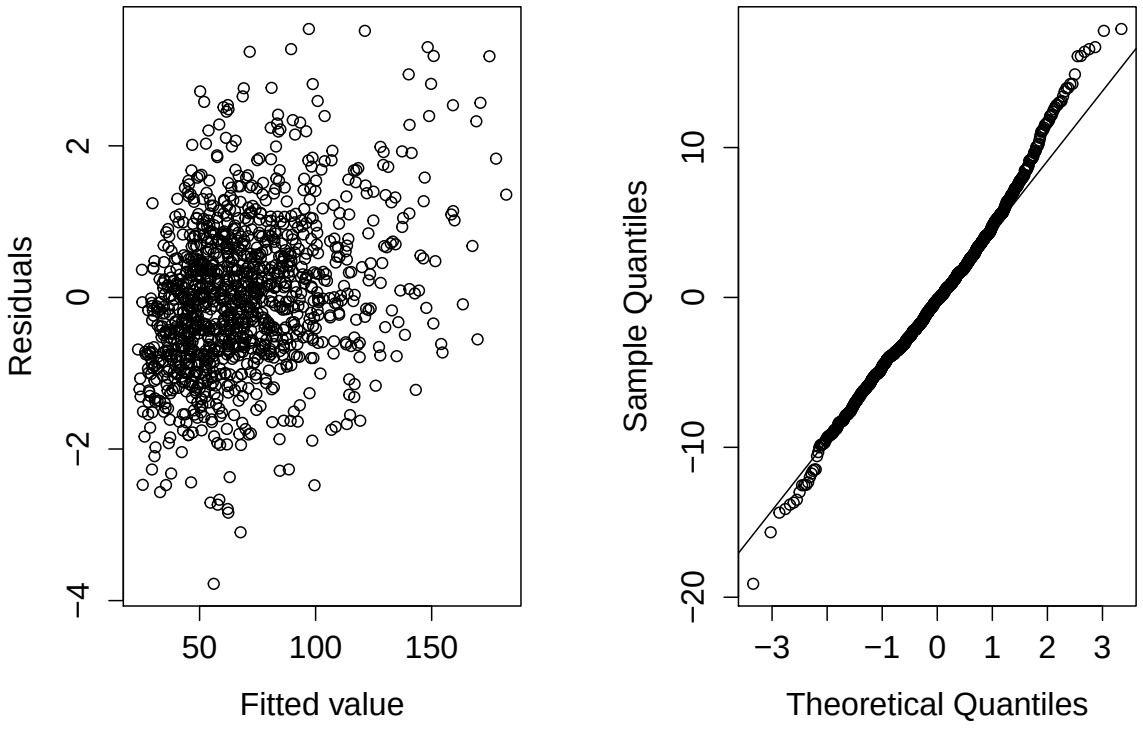


TABLE 3.5: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for mean PM_{10} with a Leroux random effect.

Variable	**Median**	**2.5%**	**97.5%**	**n.sample**	**% accetp**
(Intercept)	1.1235	0.4903	1.7678	30000	59.5
PM_{10}	0.0281	0.0158	0.0391	30000	59.5
Logprice	-0.1852	-0.2357	-0.1358	30000	59.5
JSA	0.0715	0.0641	0.0790	30000	59.5
τ^2	0.1013	0.0886	0.1156	30000	100.0
ρ	0.7800	0.6500	0.8800	30000	57.5

TABLE 3.6: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting a Poisson model for max PM_{10} with a Leroux random effect.

Variable	**Median**	**2.5%**	**97.5%**	**n.sample**	**% accetp**
(Intercept)	1.4140	0.7338	2.0356	30000	59.2
PM_{10}	0.0209	0.0111	0.0311	30000	59.2
Logprice	-0.2041	-0.2533	-0.1476	30000	59.2
JSA	0.0720	0.0642	0.0803	30000	59.2
τ^2	0.1031	0.0904	0.1171	30000	100.0
ρ	0.8000	0.6800	0.9000	30000	54.1

TABLE 3.7: Relative risk for a $1.872\mu gm^{-3}$ increase of PM_{10} from the eight models based on mean PM_{10} data in each IG.

Models	**Relative risk**	**95%credible interval**	**DIC**
G.bymCAR	1.055	(1.028, 1.082)	-734.132
G.iarCAR	1.055	(1.027, 1.082)	-232.047
G.lerouxCAR	1.055	(1.029, 1.082)	-259.467
G.cressieCAR	1.056	(1.030, 1.081)	-254.396
P.bymCAR	1.053	(1.032, 1.081)	9224.436
P.iarCAR	1.051	(1.026, 1.080)	9263.939
P.lerouxCAR	1.054	(1.030, 1.076)	9251.504
P.cressieCAR	1.053	(1.029, 1.077)	9257.276

TABLE 3.8: Relative risk for a $1.872\mu gm^{-3}$ increase of PM_{10} for eight models based on maximum PM_{10} data in each IG.

Models	**Relative risk**	**95%credible interval**	**DIC**
G.bymCAR	1.037	(1.016, 1.058)	-920.583
G.iarCAR	1.037	(1.015, 1.058)	-232.890
G.lerouxCAR	1.038	(1.018, 1.059)	-260.187
G.cressieCAR	1.038	(1.017, 1.059)	-254.859
P.bymCAR	1.035	(1.013, 1.055)	9226.393
P.iarCAR	1.035	(1.015, 1.058)	9259.560
P.lerouxCAR	1.040	(1.021, 1.060)	9250.586
P.cressieCAR	1.038	(1.019, 1.056)	9257.458

the relative risk of mean PM_{10} estimates from all models are very similar, and none of their 95% credible intervals contains the neutral value 1. This indicates that mean PM_{10} has a significant influence on respiratory disease, and the relative risk is robust. The relative risks of mean PM_{10} in Table 3.7 indicate that with a $1.872\mu gm^{-3}$ increase of mean PM_{10} concentrations, respiratory hospital admissions are likely to increase by about 5.4% (ranging from 5.1% to 5.6%). On the other hand, Table 3.8 shows that with a $1.872\mu gm^{-3}$ increase of peak PM_{10} concentration, the hospital admissions related to respiratory disease in each IG will increase about 3.7% (ranging from 3.5% to 4.0%). Therefore, both the mean and maximum spatial metrics of PM_{10} in each IG are strongly associated with the hospital admissions related to respiratory diseases.

3.6 Discussion

The effects of long-term air pollution exposure on public health in Scotland have been investigated. It was found that significant excess relative risks of respiratory hospital

admissions were associated with long-term exposures to NO_2 or PM_{10} across IGs in mainland Scotland.

For NO_2, when the spatial mean NO_2 concentration over the grid squares lying within each IG was used to represent pollution concentration, the relative risk of NO_2 is not significant. In contrast, the peak NO_2 concentrations are associated with respiratory diseases such that with a $6.84\mu gm^{-3}$ increase of peak NO_2 concentration, the hospital admissions related to respiratory disease in each IG will increase about 2.6% (ranging from 2.1% to 3.4%).

Both the spatial mean and maximum spatial metrics of PM_{10} in each IG are associated with respiratory diseases. With a $1.872\mu gm^{-3}$ increase of mean PM_{10} concentration, hospital admissions related to respiratory disease in each IG will increase about 5.4% (ranging from 5.1% to 5.6%). This value is higher than the 3.7% (ranging from 3.5% to 4.0%) obtained by using the spatial maximum PM_{10} metric.

My findings about the adverse effects of NO_2 and PM_{10} are broadly consistent with those from other recent studies. For example, Lee et al. [76] reported a relative risk of 1.04 - 1.12 for a $8\mu gm^{-3}$ increase in NO_2 concentrations and a relative risk of 1.06 - 1.10 for a $1.7\mu gm^{-3}$ increase in PM_{10} concentrations when they investigated the relationship between long-term exposures to NO_2, PM_{10} and respiratory hospital admissions in Lothian and Greater Glasgow. Belanger et al. [8] reported that exposure to indoor NO_2 at levels well below the Environmental Protection Agency outdoor standard (53 ppb) is associated with respiratory symptoms among children with asthma in multifamily housing. Kattan et al. [66] also found that higher levels of indoor NO_2 are associated with increased asthma symptoms in nonatopic children and decreased peak flows. Oftedal et al. [95] shown statistically significant respiratory health effects of exposure to NO_2. The adverse effects of PM_{10} on respiratory diseases were also reported in other researches (e.g. Lee [74], Thishan Dharshana and Coowanitwong [126] and Wiwanitkit [140]).

Chapter 4

Estimating the long-term health effects of air pollution by fusing modelled and measured pollution data

4.1 Introduction

In the study of air pollution health effects, one key problem is estimating spatially representative pollution concentrations using two main sources of data: measured data from a sparse network of monitors and modelled concentrations on a regular grid from an atmospheric dispersion model, such as those produced by AEA [1]. The latter provide complete spatial coverage of the study region but are known to contain biases (Berrocal et al. [11]). Geostatistical Kriging has been used to spatially align the monitored pollution data to the disease counts (Elliott et al. [41] and Janes et al. [62]), while simple averaging is used to correct the spatial misalignment of the modelled concentrations (Maheswaran et al. [84]; Lee et al. [76] and Warren et al. [136]). Recently, Vinikoor-Imler et al. [134], Vinikoor-Imler et al. [135], Sacks et al. [111] and Warren et al. [137] have estimated pollution using both monitored and modelled pollution data, by utilizing the fusion approaches proposed by Fuentes and Raftery [42], Berrocal et al. [11] or McMillan et al. [87].

There are two main streams of literature on fusing these two sources of data. The first one is to treat the monitoring data as the true data and use model output data as a covariate to predict point measurements. For example, Berrocal et al. [11] modelled the relationship between observations and numerical model output by taking the numerical output as data via a linear regression with spatially varying coefficients assumed to arise from correlated spatial Gaussian processes. Bruno et al. [21] calibrated radar measurements via rain gauge data, by treating rain gauges as the reference measures. Pannullo et al. [97] proposed a geostatistical fusion model that regressed combined NO_2 concentrations from both automatic monitors and diffusion tubes against modelled NO_2 concentrations from an atmospheric dispersion model to improve the prediction of NO_2 across West Central Scotland.

An alternative fusion approach is to assume an underlying unknown ground truth process which is linked separately to monitoring data and model output. For example, both Fuentes and Raftery [42] and Wikle and Berliner [138] assumed that there exists an underlying unobserved spatial process driving observational data and the numerical model output, while the former specified the true process at the point level and the latter modeled the true process at areal unit scale. The observations are related to the unobserved process via a measurement error model, while the numerical model data are via a linear model that accounts for potential bias in the model output. McMillan et al. [87] also assumed a true spatial process related to both observational data and numerical model output, however, they specified the underlying process at the block level rather than the point level. Sahu et al. [113] investigated space-time wet deposition patterns over eastern USA by developing a data fusion approach using a measurement error specification to combine gridded CMAQ output (Community Multi-scale Air Quality) and point level monitoring data. The model components have been linked using latent processes in a Bayesian hierarchical framework.

These fusion models can be classified into spatial modelling and spatio-temporal modelling according to their ability to handle spatial only data or spatio-temporal data. The former addresses data without a time dimension, with examples being Fuentes and Raftery [42], Berrocal et al. [11] and Pannullo et al. [97]. In contrast, a spatio-temporal model is used to accommodate data collected over time, with examples being Shaddick and Wakefield [117], Berrocal et al. [11], McMillan et al. [87] and Lawson et al. [72]. These fusion models usually assume spatial autocorrelation in the data.

In the previous chapter, evidence of the effects of air pollution on health has been found. The relationship between air pollution and health effect was investigated by using a benchmark method of converting modelled pollution data (DEFRA) into the small-area scale on which the disease data were collected. However, the results are only based on DEFRA data without using the measured observations, and the latter are known to be more reliable. The fusion of the measured data and DEFRA data is expected to provide improved predictions of areal level pollution concentrations.

This chapter proposes a two-stage approach to investigate the health effects of air pollution, with inference in a Bayesian setting based on McMC simulation. The first stage is a novel statistical fusion model that regresses the monitored and modelled pollution concentrations at the point-level, then makes point-level predictions of pollution across my study region, and finally aggregates these point-level predictions to the areal level required to align with the disease counts. The second stage regresses these areal level pollution summaries to the disease counts, allowing for the spatio-temporal autocorrelation in the data.

I develop my methodology for a new study investigating the long-term effects of NO_2 concentrations on respiratory disease in Scotland, UK. There have been few previous epidemiological studies of this type in Scotland, for example, only Prescott et al. [104], Carder et al. [22] and Willocks et al. [139] have investigated the association between short-term exposure to air pollution and ill health, while only Lee et al. [76] and Lee [74] have attempted to quantify the long-term effects using an ecological spatio-temporal design. The study presented in this chapter is one of the most comprehensive investigations into the effects of NO_2 concentrations on health in Scotland, as my study region is all of mainland Scotland for the five year period spanning from 2007 to 2011. In conducting this study I compare my proposed modelling approach with the simpler approach of using only the DEFRA concentrations, which allows us to assess the validity of using the latter in such ecological studies. I also consider whether the average (spatial mean) or the peak (spatial maximum) NO_2 concentration across each areal unit is an appropriate measure of exposure. The remainder of this chapter is organised as follows. In Section 4.2 I describe the study background and present some exploratory analysis, while Section 4.3 proposes my new integrated pollution and health model. The results of my study are presented in Section 4.4, while Section 4.5 provides a concluding discussion.

4.2 Background

4.2.1 Data description

As described in the previous chapter, my focus remains on mainland Scotland and the disease data are still the hospital admissions related to respiratory diseases. However, I considered 5 year disease data from 2007-2011 rather than just a single year. The disease count for area k in year t is denoted by Y_{kt}, so the set of values for all n IGs in year t is denoted by $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{nt})$. As was mentioned in chapter 3, the number of admissions in an IG depends on its population size and demographic structure, so I adjust for this by computing the expected numbers of admissions for each area in each year. These expected disease counts are denoted by E_{kt}, and the standardized incidence ratio (SIR) is given by $SIR_{kt} = Y_{kt}/E_{kt}$. Recall a spatial map of SIR for 2011 which is also shown in the top left panel of Figure 4.1. The figure shows that the majority of the high risk IGs are in the major cities of Glasgow and Edinburgh, which are the set of small densely populated IGs in the lower middle part of the country. This pattern in risk is largely driven by the geographical patterning in socio-economic deprivation, which needs to be controlled for in the model. The summary of SIR from 2007-2010 can be seen in Figure 4.2. In this chapter I still use the same proxy measures of deprivation as in chapter 3, namely the percentage of people living in each IG who are in receipt of Job Seekers Allowance (JSA), and the median property price in an area. The percentage of people in receipt of JSA in an IG ranges between 0.05% and 15.3% with a median value of 2.7%, while the median property price in an IG ranges between £22,800 and £500,000, with a median value of £125,000.

The pollutant considered in this chapter is NO_2, whose health effects have been demonstrated in the previous chapter and the existing literature, such as Ehrlich et al. [40], Tunnicliffe et al. [132], Lee et al. [76]. I use data on annual mean concentrations between 2006 to 2010 in this study rather than 2007 to 2011, to ensure that the NO_2 exposure occurred before the hospital admissions. I obtained two types of NO_2 data for my study, measured concentrations at a small number of locations and DEFRA data. The measured data are collected from two different devices, automatic monitors and diffusion tubes, and both data sets can be freely obtained from Air Quality in Scotland (*http://www.scottishairquality.co.uk/*). The data locations have been classified as either

FIGURE 4.1: Summary of the data. Top left is the SIR for respiratory disease in Scotland in 2011, top right is the modelled annual average NO_2 concentrations in 2010 (μgm^{-3}), bottom left is the locations of the measured NO_2 data (▲ for monitoring sites, + for tube sites), and bottom right shows Scotland partitioned into urban (black) and rural areas (grey).

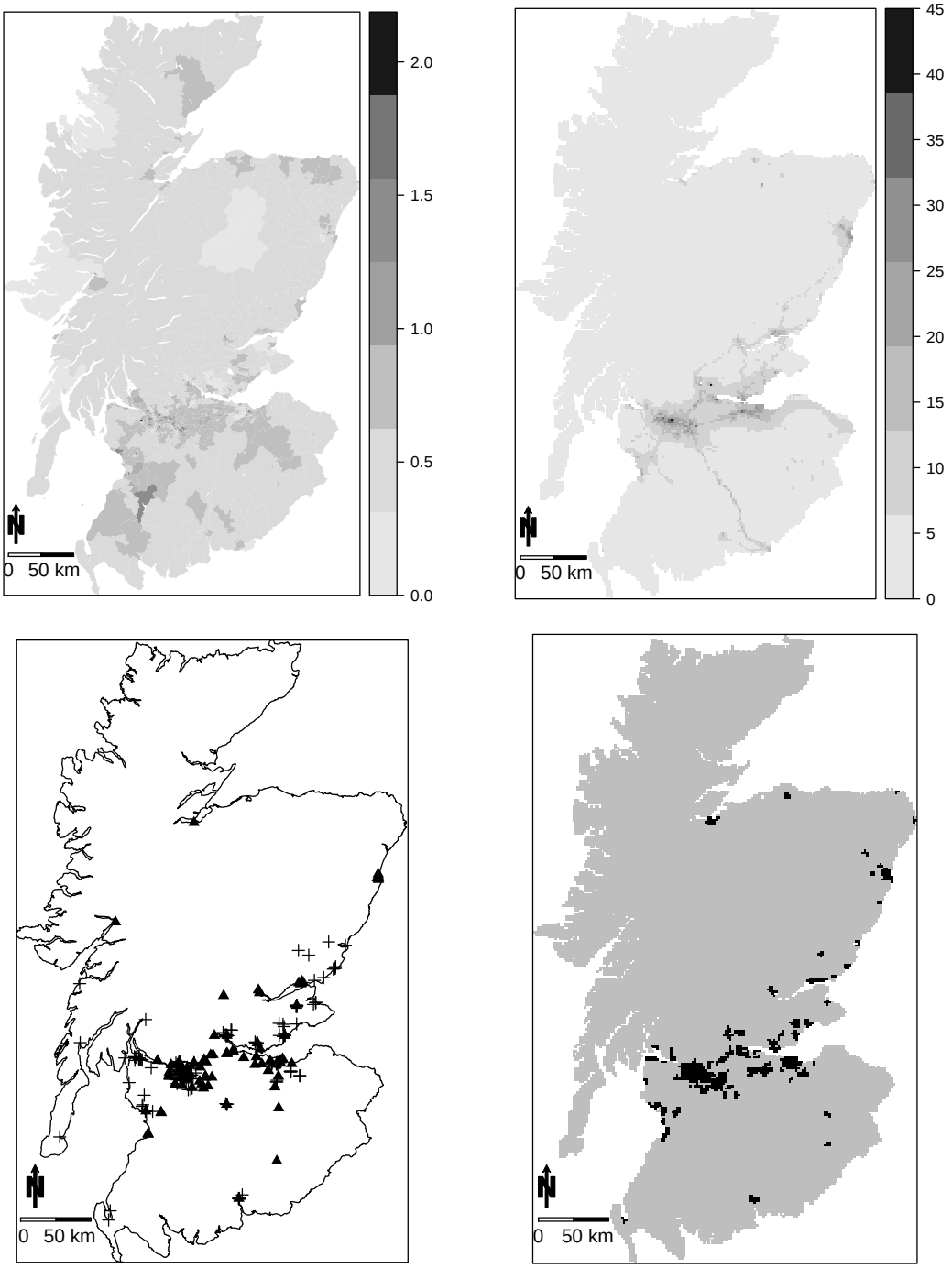
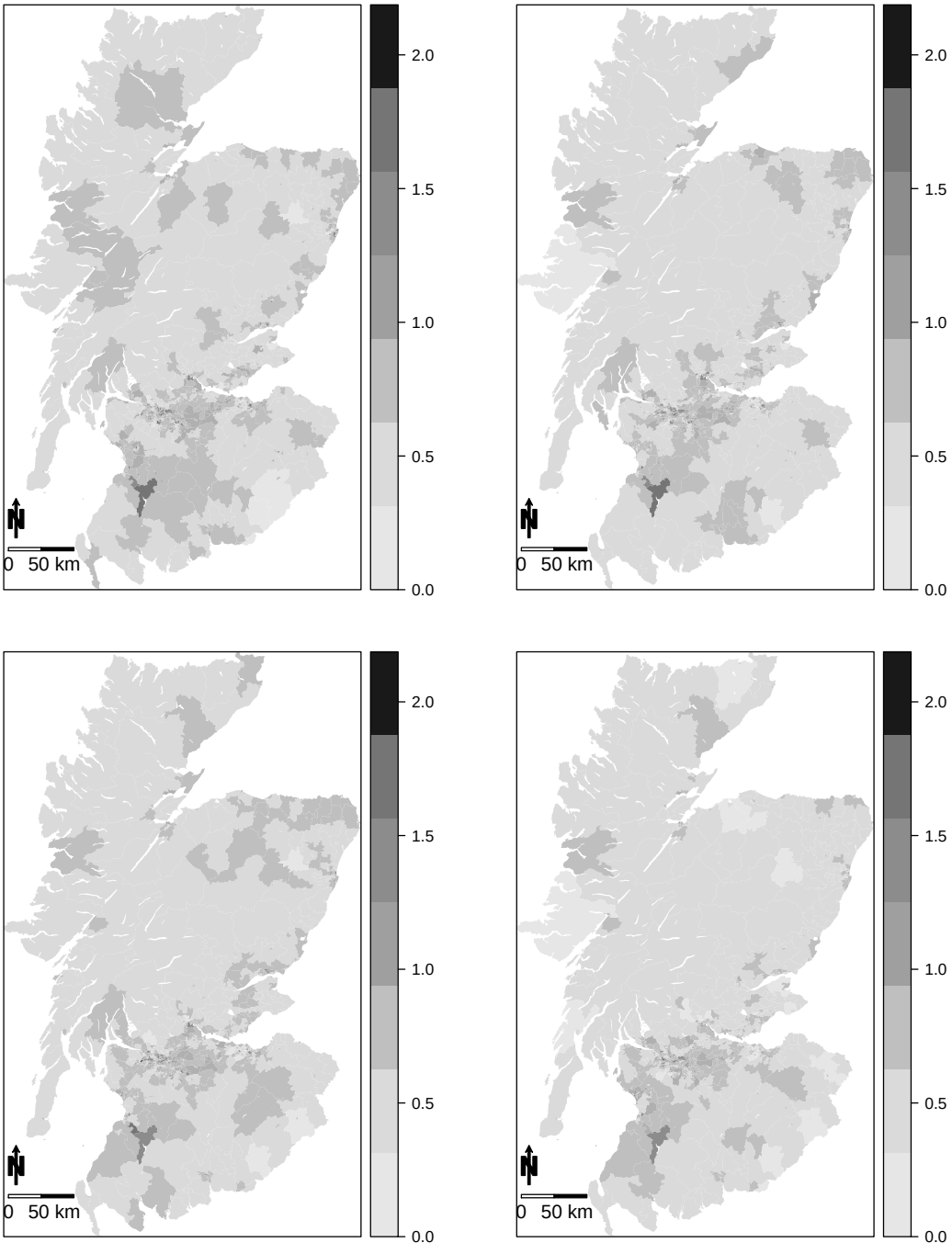


FIGURE 4.2: Summary of SIR. Top left and right are the SIR for respiratory disease in Scotland in 2007 and 2008, respectively, while the bottom left and right are for 2009 and 2010, respectively.



urban background, kerbside, roadside or rural, and a summary of the observed data is shown in Table 4.1. As might be expected, the pollution levels recorded at urban locations are higher than those at rural locations, and the closer the monitoring stations are to a main road, the higher the NO_2 concentrations are. The locations of the measured data are presented in Figure 4.1, which shows that they provide poor spatial coverage of Scotland as the major cities are well represented but the rest of the study region contains hardly any monitors. Therefore standard geostatistical prediction methods may not be appropriate here, due to the large distances between data locations and potential prediction locations.

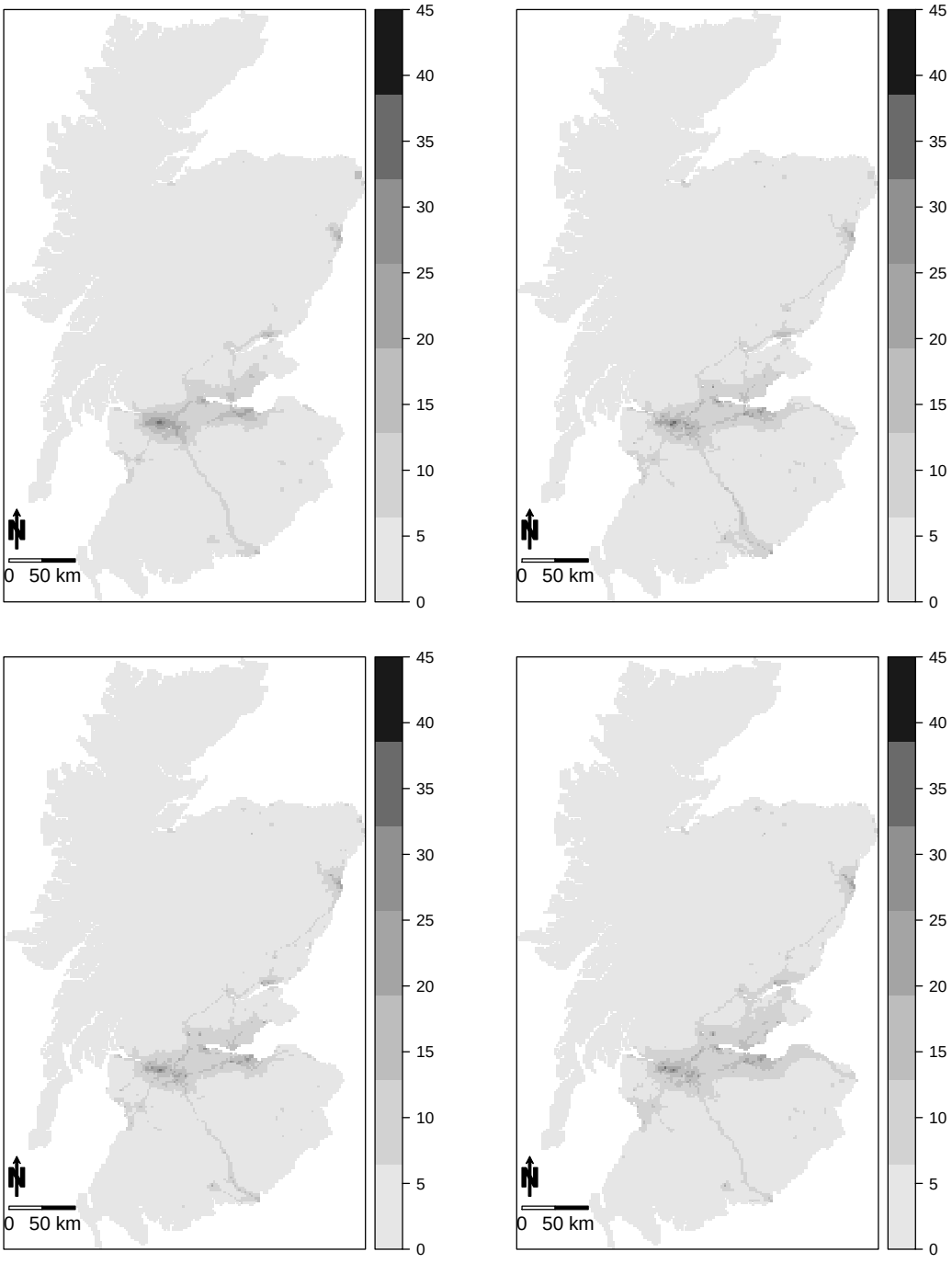
Therefore, the DEFRA data are also used in this chapter as they have complete spatial coverage of Scotland. The fusion of both measured concentrations and DEFRA data is expected to improve the prediction of exposure. The DEFRA data for 2010 is shown in the top right panel of Figure 4.1, which exhibits a similar pattern with the spatial map of SIR for 2011, with high values in the lower middle part of the country. The summary of modelled NO_2 from 2006-2009 can be seen in Figure 4.3 As temperature can affect air circulation and thus the spatial distribution of air pollution, I consider it as a covariate in my proposed pollution model outlined in Section 4.3. Temperature data are available as annual averages across Scotland at the 5km resolution from the Met Office (*http://www.metoffice.gov.uk/*), and exhibit a general north-south trend as expected.

TABLE 4.1: Summary of the measured NO_2 data by site type and year: the numbers within the round brackets represent the number of sites in the form (automatic monitors, diffusion tubes), while those within square brackets indicate their corresponding mean concentrations (μgm^{-3}).

Site type \ Year	**2006**	**2007**	**2008**	**2009**	**2010**
Urban Background	(3, 29)	(3, 29)	(6, 29)	(6, 29)	(6, 29)
	[27.3, 18.8]	[26.3, 18.4]	[27.0, 18.8]	[26.3, 20.1]	[26.0, 21.1]
Kerbside	(1, 54)	(4, 54)	(4, 54)	(3, 55)	(5, 55)
	[68.0, 31.5]	[64.0, 33.5]	[65.5, 31.2]	[67.3, 30.7]	[59.0, 32.4]
Roadside	(11, 94)	(15, 94)	(25, 95)	(30, 99)	(34, 99)
	[43.8, 33.4]	[42.4, 34.4]	[36.9, 34.4]	[36.2, 33.2]	[38.2, 34.8]
Rural	(3, 0)	(3, 0)	(3, 0)	(3, 0)	(3, 0)
	[8.0, NA]	[8.00, NA]	[8.33, NA]	[7.33, NA]	[9.33, NA]

As previously discussed the measured pollution data are classified according to their local environment, such as roadside, urban background or rural. This is likely to be an important covariate in the model, and thus I have to choose the local environment

FIGURE 4.3: Summary of modelled NO_2. Top left and right are the modelled NO_2 for respiratory disease in Scotland in 2006 and 2007, respectively, while the bottom left and right are for 2008 and 2009, respectively.



for each of my prediction locations. The set of prediction locations will be the 68,448 1km grid squares on which the modelled concentrations are computed, and hence they represent the average pollution concentrations in each 1km region. Therefore I do not specify any of the locations as roadside, as the majority of each grid square will not comprise roads (there will of course be roads in a large number of grid squares). I have to make a choice between each prediction location being urban background or rural, and for this I use the Scottish Government 8 fold Urban Rural Classification (Government [51]). The Scottish Government Urban Rural Classification provides a standard definition of rural areas in Scotland. This classification is updated every two years to incorporate the most recent Small Area Population Estimates (SAPE) produced by National Records of Scotland (NRS) and Royal Mail Postcode Address File (PAF). NRS Small Area Population Estimates (SAPE) together with information from the Royal Mail Postcode Address File (PAF) were used to classify 2010 postcode units as high or low density. This information was then used to identify areas of contiguous high density postcodes with a population of 500 or more that make up a Settlement. Details of the methodology used for the Mid-2010 Population Estimates for Settlements can be found at Mid 2010 population estimates for settlements (*http://tinyurl.com/pqvy9mw*). In my study, the 2009-2010 Urban Rural Classification is adopted, in which the data zone classification identifying urban and rural areas is based on settlement size and drive times.

The classification is available in a number of forms, including Scottish Government 2 fold Urban Rural Classification, Scottish Government 3 fold Urban Rural Classification, Scottish Government 6 fold Urban Rural Classification and Scottish Government 8 fold Urban Rural Classification. My study is based on the Scottish Government 8 fold Urban Rural Classification because its shapefile can be freely downloaded from Scottish Government (*http://tinyurl.com/oqyl36y*). The definition about this classification is shown in Table 4.2. Note that the site types of the monitoring stations include urban background, kerbside, roadside and rural. These site types in pollution model are different from the classifications in Table 4.2. Therefore, I classify any areas with settlements of over 10,000 people as urban (class 1 and 2), while the rest (class 3-8) are assumed to be rural, and this gives the map shown in the bottom right panel of Figure 4.1.

TABLE 4.2: Scottish Government 8 fold Urban Rural Classification

Classes	**Definition**
1 Large Urban Areas	Settlements of over 125,000 people.
2 Other Urban Areas	Settlements of 10,000 to 125,000 people.
3 Accessible small Towns	Settlements of between 3,000 and 10,000 people and within 30 minutes drive of a settlement of 10,000 or more.
4 Remote Small Towns	Settlements of between 3,000 and 10,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more.
5 Very Remote Small Towns	settlements of between 3,000 and 10,000 people and with a drive time of over 60 minutes to a settlement of 10,000 or more.
6 Accessible Rural	Areas with a population of less than 3,000 people, and within a 30 minute drive time of a settlement of 10,000 or more.
7 Remote Rural	Areas with a population of less than 3,000 people, and with a drive time of over 30 minutes to a settlement of 10,000 or more.
8 Very Remote Rural	areas with a population of less than 3,000 people, and with a drive time of over 60 minutes to a settlement of 10,000 or more.

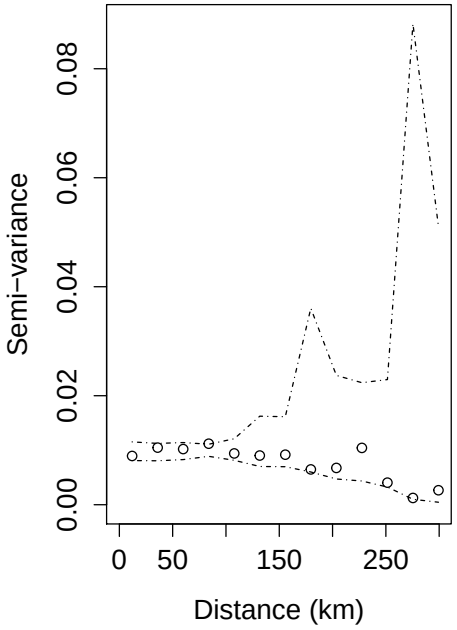
4.2.2 Exploratory analysis

I now present an exploratory analysis of the measured pollution data to inform my modelling approach proposed in Section 4.3, which aims to quantify the level of residual spatial autocorrelation remaining in these data after accounting for the known covariates. I model the measured NO_2 concentrations on the natural logarithm scale, as they are non-negative and skewed to the right and apply a simple geostatistical model to these transformed data for each year separately, where the covariates include the DEFRA concentrations (each monitoring site or diffusion tube is assigned the closest gridded DEFRA concentration) on the natural logarithm scale, the site type (e.g. roadside, rural, etc) and temperature. The geostatistical model I fit has the form

$$\boldsymbol{X} \sim N(\boldsymbol{Z\beta}, \Sigma = \sigma^2 \exp(-\boldsymbol{D}/\lambda) + \tau^2 \boldsymbol{I}), \tag{4.1}$$

where $\boldsymbol{X}$ is the vector of measured NO_2 concentrations (from the automatic monitors and diffusion tubes) for a single year. The covariates are contained in the matrix $\boldsymbol{Z}$, while $\boldsymbol{\beta}$ are the associated regression parameters. The covariance matrix is given by an exponential correlation function of distance, where $\boldsymbol{D}$ is the Euclidean distance matrix

FIGURE 4.4: The empirical semi-variogram of the residuals from the geostatistical model for 2010 (circles), with 95% Monte Carlo simulation envelopes (dashed lines).



between the data locations, σ^2 represents the partial sill, τ^2 is the nugget effect and λ is the spatial range parameter.

The model is fitted in the *geoR* (*http://www.r-project.org*) software in R, with inference based on maximum likelihood. The results show that the presence of residual spatial autocorrelation after accounting for the covariates is uncertain, as both the partial sill parameters (ranging between 0.059 and 0.083 for the five years of data) and the range parameters (ranging between 0.078 km and 0.924 km for the five years of data) are very small, and the empirical semi-variogram analysis suggests there is no or very weak residual spatial autocorrelation remaining, as the empirical semi-variogram are inside or right on the border of the Monte Carlo envelopes at all distances (see e.g. Figure 4.4 for 2010, and the semi-variogram plots for the other years are similar and are not shown here). This suggests that the available covariates, including the DEFRA concentrations (which themselves are spatially autocorrelated as shown in Figure 4.1) have captured the majority of the spatial structure in these data, and that including an additional set of spatially autocorrelated random effects is likely unnecessary.

4.2.3 Spatio-temporal pollution modelling

As described in Section 4.2.2 the pollution data contain very weak spatial autocorrelation after accounting for the covariates, and thus the spatio-temporal model proposed for the pollution data in Section 4.3 does not account for residual spatial autocorrelation. However, to assess the validity of this modelling approach I compare my proposed model against the spatio-temporal pollution model proposed by Sahu et al. [112], hereafter referred to as **SGH** which does allow for residual spatial autocorrelation. The model can be implemented using the R package *spTimer* and has the general form:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{O}_t + \boldsymbol{\epsilon}_t \quad t = 1, ..., T, \\ \boldsymbol{O}_t &= \rho \boldsymbol{O}_{t-1} + \boldsymbol{Z}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t \quad t = 2, ..., T, \end{aligned} \tag{4.2}$$

where $\boldsymbol{X}_t$ denotes the vector of measured pollution data in year t. These noisy data are modelled as a linear combination of the true values $\boldsymbol{O}_t$ and independent (white noise) errors $\boldsymbol{\epsilon}_t$. The true values are modelled with a first order autoregressive component ($\rho \boldsymbol{O}_{t-1}$), a regression component ($\boldsymbol{Z}_t \boldsymbol{\beta}$, where $\boldsymbol{Z}_t$ is the tth row of $\boldsymbol{Z}$ in Model (4.1)) and a spatial autocorrelation component $\boldsymbol{\eta}_t$. The latter is modelled independently for each time period, and is given a multivariate Gaussian prior with mean zero and an exponential correlation matrix, identical to Model (4.1).

4.3 Methodology

There are two main types of statistical fusion models developed in the literature, with the first being a regression calibration approach which regresses the measured data against the modelled concentrations via a spatially varying linear regression (see e.g. Berrocal et al. [11]; Berrocal et al. [13] and Berrocal et al. [14]). The second approach is to assume an underlying unknown ground truth process, which is informed separately by the monitoring data and model output (see e.g. Fuentes and Raftery [42]; Wikle and Berliner [138] and McMillan et al. [87]).

In this section, I propose an integrated model for estimating the long-term health effects of air pollution, that fuses DEFRA concentrations and measured pollution data to provide improved predictions of areal level pollution concentrations. As has been mentioned in Section 4.1, most of the existing epidemiological studies have used each of these data sources in isolation to estimate air pollution concentrations at the areal unit level, while only a few papers published recently attempted to examine the effects of air pollution on health by using fused estimates of monitored and modelled pollution data. Therefore, the present study will contribute to the extension of this literature which uses either only the measured pollution data (e.g. Janes et al. [62] and Young et al. [143]) or the modelled pollution data (e.g. Maheswaran et al. [84] and Lee et al. [76]) to estimate areal level pollution summaries. I propose a two-stage modelling approach to achieve this goal, the first stage of which is a spatio-temporal model that produces posterior predictive distributions for pollution concentrations at the 1 km resolution in Scotland, then an aggregation step to address the different spatial supports of the pollution and disease data. The second stage estimates the health impact of air pollution using the spatially aggregated pollution summaries.

4.3.1 Stage 1 - air pollution model

I propose a Bayesian space-time linear regression model for relating the measured concentrations to the modelled concentrations, whilst allowing for additional covariate information such as site type (e.g. roadside, rural, etc) and temperature. My model allows for temporal autocorrelation in the model parameters in adjacent years, because annual average concentrations are unlikely to change greatly from one year to the next. Conversely, I do not assume the measured concentrations are spatially autocorrelated after accounting for the covariate effects, because the exploratory analysis in Section 4.2.2 provides little evidence for the presence of such autocorrelation. Let $\boldsymbol{X}_t = (X_t(\boldsymbol{s}_1), ..., X_t(\boldsymbol{s}_{n_t}))$ denote the vector of n_t measured NO_2 concentrations (on the natural log scale) at sites $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_t})$ in year t, where $t = 1, 2, ..., T$. These measured concentrations are related to an $n_t \times p$ design matrix of covariates $\boldsymbol{Z}_t$ (including the modelled concentrations on the natural log scale), and the full model I propose is given by:

$$
\begin{aligned}
\boldsymbol{X}_t &\sim \mathrm{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma_t^2\boldsymbol{I}_t) \quad t = 1, ..., T, \\
\boldsymbol{\beta}_t &\sim \mathrm{N}\left(\boldsymbol{\beta} + \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}), \tau^2\boldsymbol{V}\right) \quad t = 2, ..., T, \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}(\boldsymbol{\beta}, \tau^2\boldsymbol{V}), \\
\boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0}, 1000\boldsymbol{V}), \\
\boldsymbol{V} &\sim \text{Inverse-Wishart}(\nu = p, \boldsymbol{\Psi} = 100\boldsymbol{I}_{p\times p}), \\
\ln(\sigma_t^2) &\sim \mathrm{N}(\ln(\sigma_{t-1}^2), \sigma^2) \quad t = 2, ..., T, \\
f(\ln(\sigma_1^2)) &\propto 1, \\
\kappa &\sim \text{Uniform}[0, 1], \\
\tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001).
\end{aligned} \tag{4.3}
$$

The measured pollution data in year t are modelled by a linear regression model with mean $\boldsymbol{Z}_t\boldsymbol{\beta}_t$ and variance $\sigma_t^2\boldsymbol{I}_t$, where $\boldsymbol{I}_t$ is an $n_t \times n_t$ identity matrix. The $p \times 1$ vector of regression parameters in the mean model $\boldsymbol{\beta}_t$ is assumed to be temporally autocorrelated, following a centred multivariate first order autoregressive process. The extent of this temporal dependence is captured by a commonly used global autoregressive parameter κ, which is assigned a uniform prior on the unit interval $[0, 1]$. If $\kappa = 0$, $\boldsymbol{\beta}_t$ is estimated independently for each year and is smoothed towards an overall mean value for all years $\boldsymbol{\beta}$, while if $\kappa = 1$, $\boldsymbol{\beta}_t$ is temporally autocorrelated with $\boldsymbol{\beta}_{t-1}$. The covariance matrix $\boldsymbol{V}$ captures the potential correlations among the elements of each $\boldsymbol{\beta}_t$, and these correlations are assumed to be constant for all years. The observation variance σ_t^2 is also assumed to be temporally autocorrelated via a first order random walk prior which is a simple and computationally efficient approach to model temporal autocorrelation, and as it must be non-negative, the log scale is used. Finally, I choose weakly informative conjugate prior distributions for $(\boldsymbol{V}, \sigma^2, \tau^2)$ by assuming them to be Inverse-Wishart and Inverse-gamma distributed respectively, where for the former $\nu = p$ and $\boldsymbol{\Psi} = 100\boldsymbol{I}_{p\times p}$ as this was used by Lawson et al. [72] as well. Inference for the collection of model parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T, \boldsymbol{\beta}, \boldsymbol{V}, \kappa, \sigma_1^2, \ldots, \sigma_T^2, \tau^2, \sigma^2)$ is based on McMC simulation, using both Gibbs sampling and Metropolis-Hastings steps. The posteriors of $\boldsymbol{\Theta}$ are achieved as follows.

Firstly, the likelihood of Model (4.3) is given by,

$$\begin{aligned} f(\boldsymbol{X}_1, ..., \boldsymbol{X}_T) &= \prod_{t=1}^{T} f(\boldsymbol{X}_t) \\ &= \prod_{t=1}^{T} \mathrm{N}(\boldsymbol{X}_t \mid \boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma_t^2\boldsymbol{I}_t). \end{aligned} \tag{4.4}$$

Then the prior can be written as,

$$\begin{aligned} f(\theta) &= f(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T, \boldsymbol{\beta}, \sigma_1^2, ..., \sigma_T^2, \sigma^2, \tau^2, \kappa, \boldsymbol{V}) \\ &= f(\sigma^2)f(\tau^2)f(\kappa)f(\boldsymbol{V})f(\boldsymbol{\beta})f(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta})f(\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_T)f(\sigma_1^2)f(\sigma_2^2, ..., \sigma_T^2) \\ &= f(\sigma^2)f(\tau^2)f(\kappa)f(\boldsymbol{V})f(\boldsymbol{\beta})f(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta})\prod_{t=2}^{T} f(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1})f(\sigma_1^2)\prod_{t=2}^{T} f(\sigma_t^2 \mid \sigma_{t-1}^2, \sigma^2) \\ &= \mathrm{IG}(\sigma^2 \mid a, b)\mathrm{IG}(\tau^2 \mid a, b)\mathrm{U}(\kappa \mid 0, 1)\mathrm{IW}(\boldsymbol{V} \mid \nu, \boldsymbol{\Psi})\mathrm{N}(\boldsymbol{\beta} \mid \boldsymbol{0}, 1000\boldsymbol{V}) \\ &\quad *\mathrm{N}(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}, \tau^2\boldsymbol{V})\prod_{t=2}^{T} \mathrm{N}\left(\boldsymbol{\beta}_t \mid \boldsymbol{\beta} + \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}), \tau^2\boldsymbol{V}\right)\prod_{t=2}^{T} \mathrm{N}\left(\ln\left(\sigma_t^2\right) \mid \ln\left(\sigma_{t-1}^2\right), \sigma^2\right). \end{aligned} \tag{4.5}$$

Basing on the likelihood and prior, I get the conditional posterior distributions for all the parameters,

$$
\begin{aligned}
f(\sigma^2 \mid -) &\propto \text{IG}\left(a+\frac{1}{2}(T-1), b+\frac{1}{2}\Sigma_{t=2}^{T}(\ln(\sigma_t^2)-\ln(\sigma_{t-1}^2))^2\right) \qquad (4.6)\\
f(\tau^2 \mid -) &\propto \text{IG}(a+\frac{T*p}{2}, b+\frac{1}{2}(\boldsymbol{\beta}_1-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\beta})+\frac{1}{2}\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t-\boldsymbol{\beta}-\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}))^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\beta}-\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})))\\
f(\boldsymbol{V} \mid -) &\propto \text{Inverse-Wishart}(\nu+T+1, \boldsymbol{\Psi}+\frac{\boldsymbol{\beta}\boldsymbol{\beta}^\top}{1000}+\frac{(\boldsymbol{\beta}_1-\boldsymbol{\beta})(\boldsymbol{\beta}_1-\boldsymbol{\beta})^\top}{\tau^2}+\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t-\boldsymbol{\beta}-\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}))(\boldsymbol{\beta}_t-\boldsymbol{\beta}-\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}))^\top}{\tau^2})\\
\ln f(\sigma_1^2) \mid - &\propto -\frac{n_1}{2}\ln(\sigma_1^2)-\frac{1}{2\sigma^2}[\ln(\sigma_2^2)-\ln(\sigma_1^2)]^2-\frac{1}{2\sigma_1^2}(\boldsymbol{X}_1-\boldsymbol{Z}_1\boldsymbol{\beta}_1)^\top(\boldsymbol{X}_1-\boldsymbol{Z}_1\boldsymbol{\beta}_1)\\
\ln f(\sigma_t^2) \mid - &\propto -\frac{n_t}{2}\ln(\sigma_t^2)-\frac{1}{2\sigma^2}[\ln(\sigma_t^2)-\ln(\sigma_{t-1}^2)]^2-\frac{1}{2\sigma^2}[\ln(\sigma_{t+1}^2)-\ln(\sigma_t^2)]^2\\
&\quad -\frac{1}{2\sigma_t^2}(\boldsymbol{X}_t-\boldsymbol{Z}_t\boldsymbol{\beta}_t)^\top(\boldsymbol{X}_t-\boldsymbol{Z}_t\boldsymbol{\beta}_t), \quad t=2,...,T-1\\
\ln f(\sigma_T^2) \mid - &\propto -\frac{n_T}{2}\ln(\sigma_T^2)-\frac{1}{2\sigma^2}[\ln(\sigma_T^2)-\ln(\sigma_{T-1}^2)]^2-\frac{1}{2\sigma_T^2}(\boldsymbol{X}_T-\boldsymbol{Z}_T\boldsymbol{\beta}_T)^\top(\boldsymbol{X}_T-\boldsymbol{Z}_T\boldsymbol{\beta}_T)\\
f(\kappa \mid -) &\propto \text{N}(\mu_\kappa, \sigma_\kappa)\\
f(\boldsymbol{\beta}_1 \mid -) &\propto \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\\
f(\boldsymbol{\beta}_t \mid -) &\propto \text{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad t=2,...,T-1\\
f(\boldsymbol{\beta}_T \mid -) &\propto \text{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)\\
f(\boldsymbol{\beta} \mid -) &\propto \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
$$

where,

$$
\begin{aligned}
\mu_\kappa &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1}\left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right] \\
\sigma_\kappa^2 &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1} \\
\boldsymbol{\mu}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{Z}_1}{\sigma_1^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}\boldsymbol{\beta}+\kappa\boldsymbol{V}^{-1}(\boldsymbol{\beta}_2-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{X}_1}{\sigma_1^2}\right] \\
\boldsymbol{\Sigma}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{Z}_1}{\sigma_1^2}\right]^{-1} \\
\boldsymbol{\mu}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma_t^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}(\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})+\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}))}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{X}_t}{\sigma_t^2}\right] \quad t=2,...,T-1 \\
\boldsymbol{\Sigma}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma_t^2}\right]^{-1} \quad t=2,...,T-1 \\
\boldsymbol{\mu}_T &= \left[\frac{\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{Z}_T}{\sigma_T^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}(\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{T-1}-\boldsymbol{\beta}))}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{X}_T}{\sigma_T^2}\right] \\
\boldsymbol{\Sigma}_T &= \left[\frac{\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{Z}_T}{\sigma_T^2}\right]^{-1} \\
\boldsymbol{\mu} &= \left[\frac{\boldsymbol{V}^{-1}}{1000}+\frac{\boldsymbol{V}^{-1}(1+(1-\kappa)^2(T-1))}{\tau^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}\boldsymbol{\beta}_1+\Sigma_{t=2}^{T}(1-\kappa)\boldsymbol{V}^{-1}(\boldsymbol{\beta}_t-\kappa\boldsymbol{\beta}_{t-1})}{\tau^2}\right] \\
\boldsymbol{\Sigma} &= \left[\frac{\boldsymbol{V}^{-1}}{1000}+\frac{\boldsymbol{V}^{-1}(1+(1-\kappa)^2(T-1))}{\tau^2}\right]^{-1}
\end{aligned}
\tag{4.7}
$$

TABLE 4.3: Simplifications of the general model (4.3).

Model	**Simplification**
1A	$\kappa = 0, \boldsymbol{V} = \boldsymbol{I}, \sigma_t^2 = \sigma^2$
1B	$\kappa = 1, \boldsymbol{V} = \boldsymbol{I}, \sigma_t^2 = \sigma^2$
1C	$\boldsymbol{V} = \boldsymbol{I}, \sigma_t^2 = \sigma^2$
1D	$\sigma_t^2 = \sigma^2$
1E	The full model

Model (4.3) is very general, and I compare its performance to a number of simplifications when modelling the NO_2 data in this chapter to see if the full model complexity is necessary for my data. The simplifications I consider are outlined in Table 4.3. Model 1A is the simplest special case and assumes the elements of $\boldsymbol{\beta}_t$ are independent of each other and over time, and additionally the observation variance σ_t^2 is assumed to be constant in time. Models 1B and 1C are similar, and respectively assume $\boldsymbol{\beta}_t$ follow first order random walk and first order autoregressive processes. Model 1D allows the full generality of the mean model for $\boldsymbol{\beta}_t$, but assumes the observation variance is constant, while Model 1E is the full model given by (4.3).

The posteriors of model parameters for Model 1A to 1D are also provided as follows.
(*Model 1A*)
The model is given by:

$$\begin{aligned} \boldsymbol{X}_t &\sim \mathrm{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma^2\boldsymbol{I}) \quad t = 1, ..., T, \\ \boldsymbol{\beta}_t &\sim \mathrm{N}\left(\boldsymbol{\beta}, \tau^2\boldsymbol{I}\right) \quad t = 1, ..., T, \\ \boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{0}, 1000\boldsymbol{I}), \\ \tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001). \end{aligned} \tag{4.8}$$

The posterior distributions for all the parameters are given by:

$$
\begin{aligned}
f(\sigma^2 \mid -) &\propto \text{IG}\left(a + \frac{1}{2}\Sigma_{t=1}^{T} n_t, b + \frac{1}{2}\Sigma_{t=1}^{T}(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)^\top(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)\right) \qquad (4.9)\\
f(\tau^2 \mid -) &\propto \text{IG}\left(a + \frac{T * p}{2}, b + \frac{1}{2}\Sigma_{t=1}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\beta})^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta})\right)\\
f(\boldsymbol{\beta}_t \mid -) &\propto \text{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad t = 1, ..., T,\\
f(\boldsymbol{\beta} \mid -) &\propto \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
$$

where,

$$
\begin{aligned}
\boldsymbol{\mu}_t &= \left[\frac{\boldsymbol{I}}{\tau^2} + \frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}}{\tau^2} + \frac{\boldsymbol{Z}_t^\top \boldsymbol{X}_t}{\sigma^2}\right], \quad t = 1, ..., T, \qquad (4.10)\\
\boldsymbol{\Sigma}_t &= \left[\frac{\boldsymbol{I}}{\tau^2} + \frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma^2}\right]^{-1}, \quad t = 1, ..., T,\\
\boldsymbol{\mu} &= \left[\frac{\boldsymbol{I}}{1000} + \frac{T}{\tau^2}\boldsymbol{I}\right]^{-1}\left[\frac{\Sigma_{t=1}^{T}\boldsymbol{\beta}_t}{\tau^2}\right]\\
\boldsymbol{\Sigma} &= \left[\frac{\boldsymbol{I}}{1000} + \frac{T}{\tau^2}\boldsymbol{I}\right]^{-1}
\end{aligned}
$$

(*Model 1B*)

The model is given by:

$$
\begin{aligned}
\boldsymbol{X}_t &\sim \text{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma^2\boldsymbol{I}) \quad t = 1, ..., T, \qquad (4.11)\\
\boldsymbol{\beta}_t &\sim \text{N}\left(\boldsymbol{\beta}_{t-1}, \tau^2\boldsymbol{I}\right) \quad t = 2, ..., T,\\
\boldsymbol{\beta}_1 &\sim \text{N}(\boldsymbol{0}, 1000\boldsymbol{I}),\\
\tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001).
\end{aligned}
$$

The posterior distributions for all the parameters are given by:

$$
\begin{aligned}
f(\sigma^2 \mid -) &\propto \text{IG}\left(a + \frac{1}{2}\Sigma_{t=1}^{T} n_t, b + \frac{1}{2}\Sigma_{t=1}^{T}(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)^\top(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)\right) \quad (4.12)\\
f(\tau^2 \mid -) &\propto \text{IG}\left(a + \frac{(T-1)*p}{2}, b + \frac{1}{2}\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})\right)\\
f(\boldsymbol{\beta}_1 \mid -) &\propto \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),\\
f(\boldsymbol{\beta}_t \mid -) &\propto \text{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad t = 2, ..., T-1,\\
f(\boldsymbol{\beta}_T \mid -) &\propto \text{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)
\end{aligned}
$$

where,

$$
\begin{aligned}
\boldsymbol{\mu}_1 &= \left[(\frac{1}{\tau^2} + \frac{1}{1000})\boldsymbol{I} + \frac{\boldsymbol{Z}_1^\top\boldsymbol{Z}_1}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}_2}{\tau^2} + \frac{\boldsymbol{Z}_1^\top\boldsymbol{X}_1}{\sigma^2}\right] \quad (4.13)\\
\boldsymbol{\Sigma}_1 &= \left[(\frac{1}{\tau^2} + \frac{1}{1000})\boldsymbol{I} + \frac{\boldsymbol{Z}_1^\top\boldsymbol{Z}_1}{\sigma^2}\right]^{-1}\\
\boldsymbol{\mu}_t &= \left[\frac{2}{\tau^2}\boldsymbol{I} + \frac{\boldsymbol{Z}_t^\top\boldsymbol{Z}_t}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}_{t-1} + \boldsymbol{\beta}_{t+1}}{\tau^2} + \frac{\boldsymbol{Z}_t^\top\boldsymbol{X}_t}{\sigma^2}\right], \quad t = 2, ..., T-1\\
\boldsymbol{\Sigma}_t &= \left[\frac{2}{\tau^2}\boldsymbol{I} + \frac{\boldsymbol{Z}_t^\top\boldsymbol{Z}_t}{\sigma^2}\right]^{-1}, \quad t = 2, ..., T-1\\
\boldsymbol{\mu}_T &= \left[\frac{\boldsymbol{I}}{\tau^2} + \frac{\boldsymbol{Z}_T^\top\boldsymbol{Z}_T}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}_{T-1}}{\tau^2} + \frac{\boldsymbol{Z}_T^\top\boldsymbol{X}_T}{\sigma^2}\right]\\
\boldsymbol{\Sigma}_T &= \left[\frac{\boldsymbol{I}}{\tau^2} + \frac{\boldsymbol{Z}_T^\top\boldsymbol{Z}_T}{\sigma^2}\right]^{-1}
\end{aligned}
$$

(*Model 1C*)

The model is given by:

$$
\begin{aligned}
\boldsymbol{X}_t &\sim \text{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma^2\boldsymbol{I}) \quad t = 1, ..., T, \quad (4.14)\\
\boldsymbol{\beta}_t &\sim \text{N}\left(\boldsymbol{\beta} + \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}), \tau^2\boldsymbol{I}\right) \quad t = 2, ..., T,\\
\boldsymbol{\beta}_1 &\sim \text{N}(\boldsymbol{\beta}, \tau^2\boldsymbol{I}),\\
\boldsymbol{\beta} &\sim \text{N}(\boldsymbol{0}, 1000\boldsymbol{I}),\\
\kappa &\sim \text{Uniform}[0, 1],\\
\tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001).
\end{aligned}
$$

The posterior distributions for all the parameters are given by:

$$
\begin{aligned}
f(\sigma^2 \mid -) &\propto \text{IG}\left(a + \frac{1}{2}\Sigma_{t=1}^{T} n_t, b + \frac{1}{2}\Sigma_{t=1}^{T}(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)^\top(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)\right) \\
f(\tau^2 \mid -) &\propto \text{IG}\left(a + \frac{T*p}{2}, b + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta})^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}) + \frac{1}{2}\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))\right) \\
f(\kappa \mid -) &\propto \text{N}(\mu_\kappa, \sigma_\kappa) \\
f(\boldsymbol{\beta}_1 \mid -) &\propto \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\
f(\boldsymbol{\beta}_t \mid -) &\propto \text{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad t = 2, ..., T-1 \\
f(\boldsymbol{\beta}_T \mid -) &\propto \text{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T) \\
f(\boldsymbol{\beta} \mid -) &\propto \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
\tag{4.15}
$$

where,

$$
\begin{aligned}
\mu_\kappa &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1}\left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t}-\boldsymbol{\beta})^\top(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right] \\
\sigma_\kappa^2 &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1} \\
\boldsymbol{\mu}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top\boldsymbol{Z}_1}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_2-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})}{\tau^2}+\frac{\boldsymbol{Z}_1^\top\boldsymbol{X}_1}{\sigma^2}\right] \\
\boldsymbol{\Sigma}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top\boldsymbol{Z}_1}{\sigma^2}\right]^{-1} \\
\boldsymbol{\mu}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top\boldsymbol{Z}_t}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})+\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}+\frac{\boldsymbol{Z}_t^\top\boldsymbol{X}_t}{\sigma^2}\right] \quad t=2,...,T-1 \\
\boldsymbol{\Sigma}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top\boldsymbol{Z}_t}{\sigma^2}\right]^{-1} \quad t=2,...,T-1 \\
\boldsymbol{\mu}_T &= \left[\frac{\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top\boldsymbol{Z}_T}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{T-1}-\boldsymbol{\beta})}{\tau^2}+\frac{\boldsymbol{Z}_T^\top\boldsymbol{X}_T}{\sigma^2}\right] \\
\boldsymbol{\Sigma}_T &= \left[\frac{\boldsymbol{I}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top\boldsymbol{Z}_T}{\sigma^2}\right]^{-1} \\
\boldsymbol{\mu} &= \left[\frac{\boldsymbol{I}}{1000}+\frac{1+(1-\kappa)^2(T-1)}{\tau^2}\boldsymbol{I}\right]^{-1}\left[\frac{\boldsymbol{\beta}_1+\Sigma_{t=2}^{T}(1-\kappa)(\boldsymbol{\beta}_t-\kappa\boldsymbol{\beta}_{t-1})}{\tau^2}\right] \\
\boldsymbol{\Sigma} &= \left[\frac{\boldsymbol{I}}{1000}+\frac{1+(1-\kappa)^2(T-1)}{\tau^2}\boldsymbol{I}\right]^{-1}
\end{aligned}
\tag{4.16}
$$

(*Model 1D*)

The model is given by:

$$
\begin{aligned}
\boldsymbol{X}_t &\sim \mathrm{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma^2\boldsymbol{I}) \quad t = 1, ..., T, \\
\boldsymbol{\beta}_t &\sim \mathrm{N}\left(\boldsymbol{\beta} + \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}), \tau^2\boldsymbol{V}\right) \quad t = 2, ..., T, \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}(\boldsymbol{\beta}, \tau^2\boldsymbol{V}), \\
\boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0}, 1000\boldsymbol{V}), \\
\boldsymbol{V} &\sim \text{Inverse-Wishart}(\nu = p, \boldsymbol{\Psi} = 100\boldsymbol{I}_{p\times p}), \\
\kappa &\sim \text{Uniform}[0, 1], \\
\tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001).
\end{aligned}
\tag{4.17}
$$

The posterior distributions for all the parameters are given by:

$$
\begin{aligned}
f(\sigma^2 \mid -) &\propto \text{IG}\left(a + \frac{1}{2}\Sigma_{t=1}^{T} n_t, b + \frac{1}{2}\Sigma_{t=1}^{T}(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)^\top(\boldsymbol{X}_t - \boldsymbol{Z}_t\boldsymbol{\beta}_t)\right) \qquad (4.18)\\
f(\tau^2 \mid -) &\propto \text{IG}\left(a + \frac{T*p}{2}, b + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}) + \frac{1}{2}\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))\right)\\
f(\boldsymbol{V} \mid -) &\propto \text{Inverse-Wishart}\left(\nu + T + 1, \boldsymbol{\Psi} + \frac{\boldsymbol{\beta}\boldsymbol{\beta}^\top}{1000} + \frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta})(\boldsymbol{\beta}_1 - \boldsymbol{\beta})^\top}{\tau^2} + \frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))(\boldsymbol{\beta}_t - \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}))^\top}{\tau^2}\right)\\
f(\kappa \mid -) &\propto \text{N}(\mu_\kappa, \sigma_\kappa)\\
f(\boldsymbol{\beta}_1 \mid -) &\propto \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\\
f(\boldsymbol{\beta}_t \mid -) &\propto \text{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad t = 2, ..., T-1\\
f(\boldsymbol{\beta}_T \mid -) &\propto \text{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)\\
f(\boldsymbol{\beta} \mid -) &\propto \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
$$

where,

$$
\begin{aligned}
\mu_\kappa &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1}\left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right] \\
\sigma_\kappa^2 &= \left[\frac{\Sigma_{t=2}^{T}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta})}{\tau^2}\right]^{-1} \\
\boldsymbol{\mu}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{Z}_1}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}\boldsymbol{\beta}+\kappa\boldsymbol{V}^{-1}(\boldsymbol{\beta}_2-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{X}_1}{\sigma^2}\right] \\
\boldsymbol{\Sigma}_1 &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_1^\top \boldsymbol{Z}_1}{\sigma^2}\right]^{-1} \\
\boldsymbol{\mu}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}(\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}+\kappa\boldsymbol{\beta})+\kappa(\boldsymbol{\beta}_{t-1}-\boldsymbol{\beta}))}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{X}_t}{\sigma^2}\right] \quad t=2,...,T-1 \\
\boldsymbol{\Sigma}_t &= \left[\frac{(1+\kappa^2)\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{\sigma^2}\right]^{-1} \quad t=2,...,T-1 \\
\boldsymbol{\mu}_T &= \left[\frac{\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{Z}_T}{\sigma^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}(\boldsymbol{\beta}+\kappa(\boldsymbol{\beta}_{T-1}-\boldsymbol{\beta}))}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{X}_T}{\sigma^2}\right] \\
\boldsymbol{\Sigma}_T &= \left[\frac{\boldsymbol{V}^{-1}}{\tau^2}+\frac{\boldsymbol{Z}_T^\top \boldsymbol{Z}_T}{\sigma^2}\right]^{-1} \\
\boldsymbol{\mu} &= \left[\frac{\boldsymbol{V}^{-1}}{1000}+\frac{\boldsymbol{V}^{-1}(1+(1-\kappa)^2(T-1))}{\tau^2}\right]^{-1}\left[\frac{\boldsymbol{V}^{-1}\boldsymbol{\beta}_1+\Sigma_{t=2}^{T}(1-\kappa)\boldsymbol{V}^{-1}(\boldsymbol{\beta}_t-\kappa\boldsymbol{\beta}_{t-1})}{\tau^2}\right] \\
\boldsymbol{\Sigma} &= \left[\frac{\boldsymbol{V}^{-1}}{1000}+\frac{\boldsymbol{V}^{-1}(1+(1-\kappa)^2(T-1))}{\tau^2}\right]^{-1}
\end{aligned}
\tag{4.19}
$$

The pollution model (4.3) is used to predict the pollution concentrations at 1 km resolution across mainland Scotland, which results in 68,448 prediction locations for each of $T = 5$ time periods (years). For a single location $\mathbf{s}_*$ and time period t, predictions are made from the posterior predictive distribution $f(X_t(\mathbf{s}_*)|\mathbf{X})$, where $\mathbf{X}$ denotes the vector of measured pollution data on the natural log scale for all time periods. M predictions are made from each posterior predictive distribution via composition sampling, sampling from the distribution $\mathrm{N}(\boldsymbol{Z}_{*t}^{\top}\boldsymbol{\beta}_t, \sigma_t^2\boldsymbol{I}_t)$, using the equation

$$X_t^{(m)}(\boldsymbol{s}_*) \mid \boldsymbol{\Theta}^{(m)} \sim \mathrm{N}(\boldsymbol{Z}_{*t}^{\top}\boldsymbol{\beta}_t^{(m)}, \sigma_t^{2(m)}\boldsymbol{I}_t) \quad m = 1, \ldots, M,$$

where $^{(m)}$ denotes the mth McMC sample drawn from the posterior distribution of the model parameters and $\boldsymbol{Z}_{*t}$ is the corresponding vector of covariates for the prediction location $\mathbf{s}_*$ at time t. The posterior mean of the M exponentiated predictions (as the measured data were modelled on the natural log scale) is taken at each grid point, resulting in $Q = 68,448$ spatial point predictions $(\tilde{X}_t(\mathbf{s}_{1*}) \ldots, \tilde{X}_t(\mathbf{s}_{Q*}))$ for each of $T = 5$ time periods. The disease data relate to irregularly shaped geographical units, and are thus spatially misaligned to the point level pollution predictions. Therefore I consider two different spatial aggregation approaches here, the spatial mean and the spatial maximum value in each areal unit. Specifically, for areal unit k and time period t I consider the following two metrics:

$$\tilde{X}_{kt}^{(1)} = \frac{1}{N_k}\sum_{r\in\mathcal{A}_k}\tilde{X}_t(\mathbf{s}_{r*}) \quad \tilde{X}_{kt}^{(2)} = \max_{r\in\mathcal{A}_k}\{\tilde{X}_t(\mathbf{s}_{r*})\}, \tag{4.20}$$

where $\mathcal{A}_k$ is the set of prediction locations (centroid of each 1 km grid) that fall within the kth areal unit, while N_k is the cardinality of this set. For an areal unit without any prediction locations located, $\mathcal{A}_k$ represents the nearest prediction location from the centroid of that areal unit and $N_k = 1$. N_k can be very different due to different area for the IGs (see the top right in Figure 4.1). For example, there are 135 IGs without any prediction locations located and 364 IGs with 1, while there are 129 IGs with more than 100 prediction locations. I note that various aggregation functions for transferring spatial

data into a single metric have been discussed by researchers (see e.g. Bruno and Cocchi [20]), however, the existing literature in the context of investigating air pollution health effects uses the mean almost exclusively (e.g. Maheswaran et al. [84] and Lee et al. [76]), whereas here I investigate both metrics as it may be that peak concentrations (over space) are more correlated with disease risk than average concentrations. Note that the population are usually not uniformly spread across an areal unit, a population weighted average of the predictions might be more reasonable. However, the population distribution information is usually lacking in practice, which limits its application.

4.3.2 Stage 2 - disease model

Recall from Section 4.2 that (Y_{kt}, E_{kt}) are the observed and expected numbers of disease cases in areal unit k during time period t, and the model presented here relates the pollution metrics in equation (4.20) to these disease counts whilst accounting for other covariate factors and spatio-temporal autocorrelation. The model I use was developed by Rushworth et al. [109], and is given by:

$$\begin{aligned}
Y_{kt} \mid E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\
\ln(R_{kt}) &= \boldsymbol{b}_{kt}^T\boldsymbol{\alpha} + \tilde{X}^{(j)}_{k(t-1)}\lambda + \phi_{kt}, \\
\boldsymbol{\alpha} &\sim \text{N}\left(\mathbf{0}, 1000\boldsymbol{I}\right), \\
\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1} &\sim \text{N}\left(\gamma\boldsymbol{\phi}_{t-1}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right), t \in 2, ..., T, \\
\boldsymbol{\phi}_1 &\sim \text{N}\left(\mathbf{0}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right), \\
\lambda &\sim \text{N}\left(0, 1000\right), \\
\nu^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001), \\
\gamma, \rho &\sim \text{U}[0, 1].
\end{aligned} \tag{4.21}$$

The risk of disease in areal unit k and time period t is denoted by R_{kt}, and is modelled by three components on the log-scale. The first is a vector of covariates, $\boldsymbol{b}_{kt}$ such as measures of poverty, and $\boldsymbol{\alpha}$ are the corresponding regression parameters which are assigned a zero-mean Gaussian prior with a diagonal variance matrix and a large variance. The pollution metric used in this model is $\tilde{X}^{(j)}_{kt}$ from equation (4.20), where $j = 1, 2$ denotes

the spatial mean and spatial maximum pollution concentration respectively. The key parameter of interest in this model is λ, the increase in the log-risk of disease for a 1 unit increase in pollution, and this is assigned a weakly informative Gaussian prior with a large variance. Note that the linear effect for the exposure in model 4.21 is suggested by the scatter plots between the natural log of SIR and the exposure which are not shown here.

The final term in the model is ϕ_{kt}, which is a random effect included to allow for any spatio-temporal autocorrelation remaining in the disease counts after the covariate effects have been accounted for. Here $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{nt})$ denotes the vector of random effects for time period t, and is modelled by a multivariate first order autoregressive process with temporal autocorrelation parameter γ and variance ν^2. Spatial autocorrelation is induced into the random effects by the precision matrix, which is given by $\boldsymbol{Q}(\rho, \boldsymbol{W}) = \rho(\text{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}) + (1-\rho)\boldsymbol{I}$ and corresponds to the conditional autoregressive (CAR) prior proposed by Leroux et al. [80]. Here spatial similarity is determined by a binary $n \times n$ adjacency matrix $\boldsymbol{W}$, which is based on the contiguity structure of the n areal units. In this matrix $w_{kk'} = 1$ if areal unit k shares a border with areal unit k', otherwise $w_{kk'} = 0$, and also $w_{kk} = 0$ for all k. The level of spatial autocorrelation in the random effects is controlled by ρ, and this can be more clearly seen by re-writing the prior for $\boldsymbol{\phi}_1$ in its full conditional form $f(\phi_{k1}|\boldsymbol{\phi}_{-k1})$, where $\boldsymbol{\phi}_{-k1}$ denotes the vector of random effects for time period 1 except for ϕ_{k1}. This full conditional distribution is given by

$$\phi_{k1}|\boldsymbol{\phi}_{-k1} \sim \text{N}\left(\frac{\rho\sum_{i=1}^{n} w_{ki}\phi_{i1}}{\rho\sum_{i=1}^{n} w_{ki} + 1 - \rho}, \frac{\nu^2}{\rho\sum_{i=1}^{n} w_{ki} + 1 - \rho}\right), \tag{4.22}$$

and if $\rho = 0$ the random effects are *a-priori* independent with mean zero and a constant variance. In contrast if $\rho = 1$ the random effects are spatially autocorrelated, as the conditional expectation of ϕ_{k1} is the mean of the random effects in neighbouring areal units while the variance is inversely proportional to the number of neighbouring units. Further details about the specification of this model is given in Rushworth et al. [109]. Finally, I choose weakly informative hyperpriors for the parameters (ν^2, ρ, γ), which allows their values to be informed by the data. Inference for the collection of model parameters $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \lambda, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T, \nu^2, \gamma, \rho)$ are based on McMC simulation, using both

Gibbs sampling and Metropolis-Hastings steps, and was implemented using the *R* package *CARBayesST* which is freely available to download from *http://cran.r-project.org*.

4.4 Results

I now present the results of my study investigating the long-term effects of NO_2 concentrations on respiratory hospitalisation risk in mainland Scotland between 2007 and 2011. Section 4.4.1 presents a validation study comparing the predictive performance of a number of different pollution models, and section 4.4.2 summarises the predictions from the best performing pollution model. Section 4.4.3 presents the estimated health effects, while section 4.4.4 tests the robustness of the health associations. For all the results presented in this section, inference is achieved using McMC simulation, where the Markov chain was burnt in for 20,000 iterations and then the remaining 30,000 iterations were used for the final results.

4.4.1 Pollution model validation

In this section I compare the predictive performance of the five variants of the air pollution model (4.3) proposed here and summarised in Table 4.3 with two alternatives, the Gaussian process model (4.2) (referred to as **SGH**), and simply using the DEFRA concentrations in isolation. I also validate the use of DEFRA concentrations in pollution models by running two extra pollution models without using the DEFRA concentrations as a covariate, Model 1E and Model SGH. I measure predictive performance using a 10-fold cross validation approach, where in each run I leave out 15% of the non-rural sites as a test set (only 3 rural sites are contained in the data and removing them might cause unstable prediction), and fit each model to the remaining data and predict the pollution concentrations in the test set. I quantify model performance by computing the prediction bias, root mean square prediction error (RMSPE) and the coverage probabilities of the 95% prediction intervals. These results are presented in Table 4.4, and as previously discussed all models are fitted to the pollution data on the natural log scale.

The table shows a number of key results. Firstly, the five variants of the pollution model proposed here give almost identical results, with negligible bias, lower RMSPE than the other models considered and close to the nominal 95% coverage probabilities. Thus

my proposed model outperforms the competitors considered here, and will be used for pollution estimation in the remainder of this section. Specifically, as Model 1A is simpler than the other variants proposed here and performs comparably, I use it for predicting pollution concentrations to be used in the disease model. The comparable performances of Models 1A and 1E for our data are because the estimated error variances σ_t^2 from the latter are very similar in each year, with posterior means of 0.096, 0.096, 0.095, 0.091, 0.089 for the five years. Furthermore, the other simplification that the covariance matrix $\mathbf{V} = \mathbf{I}$ is also not unrealistic, as the off diagonal elements of this matrix estimated from 1E are much smaller (ranging between -7.3 and 6.2) than the diagonal ones (ranging between 28.4 and 48.3).

Model **SGH** has an RMSPE that is around 24% higher than those from Models 1A to 1E, despite all models having the same covariates. This is because the spatial random effects in Model **SGH** are competing with the covariates to explain the variation in the response, resulting in attenuation in the estimated covariate effects. This is observed in Table 4.5, where the regression coefficients from Model **SGH** are smaller in absolute value than the corresponding estimates from Model 1A. This results in poorer prediction because the DEFRA concentrations are naturally a better predictor of the measured concentrations than a spatial random effect. Secondly, the prediction intervals from Model **SGH** are too wide with a coverage of 100%, which is likely to be because it has much larger standard deviation parameters compared with Model 1A (the observation standard deviations are 0.30 and 0.96 for Models 1A and **SGH** respectively). Table 4.4 also shows that using the DEFRA concentrations in isolation results in poorer spatial prediction than using both sources of data, with a RMSPE of 0.86 compared with 0.31 for the models proposed here. Finally, Table 4.4 shows that the DEFRA concentrations are an important covariate in the air pollution model as they can reduce RMSPE. Specifically, Model 1E without DEFRA concentrations has an RMSPE that is around 26% higher than that from Model 1E with DEFRA concentrations, while this value is about 17% for Model **SGH**.

4.4.2 Pollution model prediction

Since Model 1A performed as well as Model 1E, I use it to make pollution predictions at the 1km resolution across mainland Scotland. As described in Section 4.3 posterior predictive mean concentrations were computed at $Q = 68,448$ prediction locations,

TABLE 4.4: Bias, root mean square prediction error and coverage probabilities from a 10 fold cross validation excercise for the models proposed in this chapter, the autoregressive Gaussian process model (**SGH**) and using only the DEFRA concentrations.

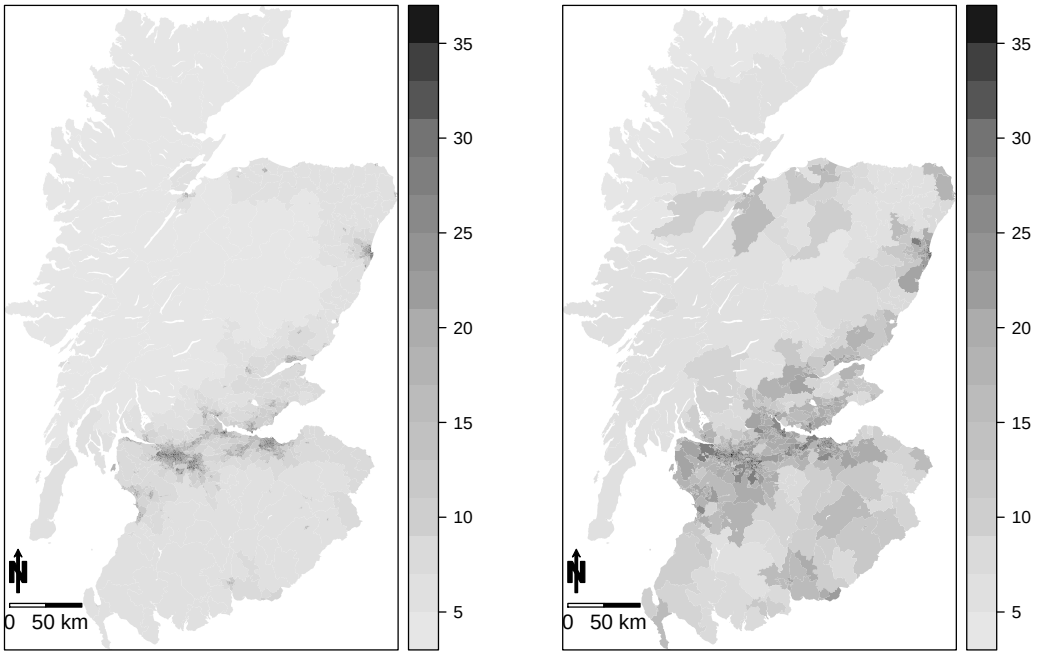
Model	**Bias**	**RMSPE**	**Coverage**
DEFRA NO_2	-0.7377	0.8648	–
1A	0.0250	0.3116	93.86%
1B	0.0249	0.3117	93.67%
1C	0.0249	0.3117	93.99%
1D	0.0250	0.3124	93.80%
1E	0.0259	0.3113	93.80%
1E without DEFRA NO_2	0.0158	0.3927	93.99%
SGH	0.0184	0.4174	100%
SGH without DEFRA NO_2	0.0210	0.4878	100%

TABLE 4.5: Posterior means for the regression parameters from Model 1A and the Gaussian process model **SGH**. The five columns ($\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5$) are the yearly regression parameter estimates from Model 1A, whlie Model **SGH** has constant regression parameters over time (final column).

Parameter	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$	$\boldsymbol{\beta}_5$	**SGH**
Kerbside	0.577	0.580	0.569	0.568	0.576	0.294
Roadside	0.592	0.597	0.594	0.587	0.595	0.304
Rural	-0.592	-0.588	-0.587	-0.590	-0.588	-0.068
DEFRA concentrations	0.375	0.541	0.549	0.516	0.475	0.142
datatype	0.154	0.158	0.145	0.139	0.144	-0.012
Temperature	0.078	0.091	0.082	0.069	0.073	0.052

and were then aggregated to the IG scale using both the spatial mean and the spatial maximum (see equation (4.20)). These areal level summaries are shown in Figure 4.5, and will be used in the disease model in the next subsection. The plots show that air pollution is highest in the most densely populated cities of Glasgow and Edinburgh, in the middle (north to south) of mainland Scotland. This pattern is similar to the spatial map of DEFRA concentrations for 2010 shown in Figure 4.1 because the latter is naturally an important predictor of the measured data. The correlations between the DEFRA and predicted pollution concentrations are high, being 0.918 for the spatial mean across an IG and 0.885 for the spatial maximum. Additionally, the spatial mean and maximum estimates from Model 1A are highly correlated, as a Pearson's correlation coefficient between the mean and maximum concentrations across an IG is 0.884. Note that the DEFRA concentrations are lower on average than the predictions from Model 1A, especially for those urban background grids. For example, a scatterplot of all model-predicted versus DEFRA values for 2010 is shown on the left of Figure 4.6, and a spatial map of the differences between model-predicted and DEFRA values on the right.

FIGURE 4.5: Spatially aggregated predicted NO_2 concentrations (μgm^{-3}) from Model 1A for 2010. The left panel shows the spatial mean concentration over each IG, while the right panel shows the spatial maximum concentration over each IG.

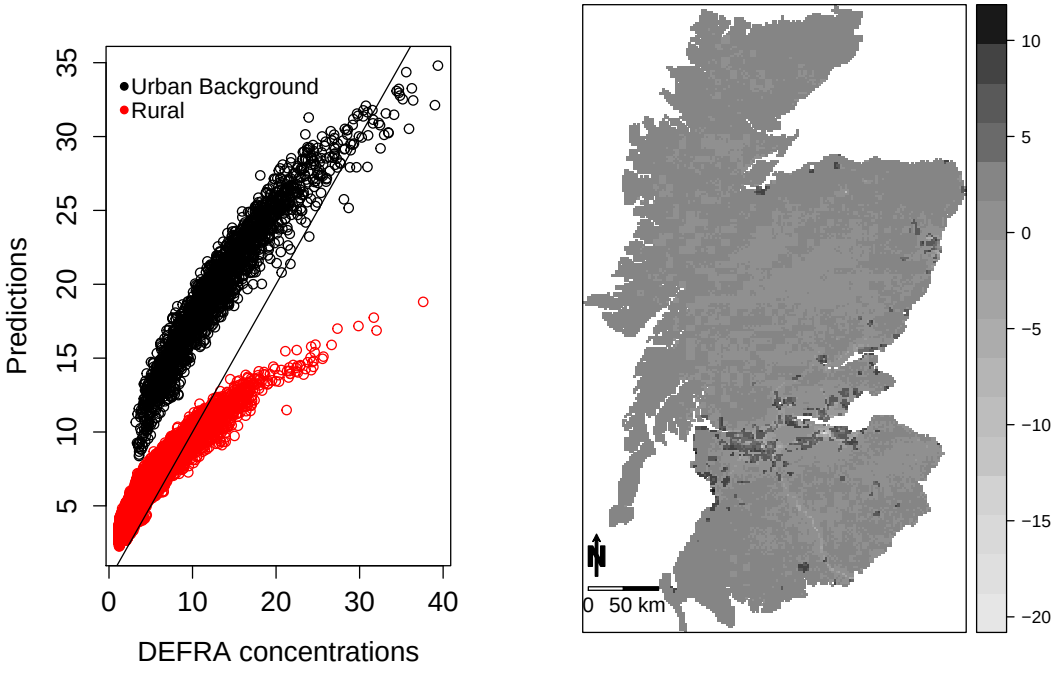


4.4.3 Disease model results

I begin by assessing the necessity of allowing for spatio-temporal autocorrelation in the disease data via the random effects in Model (4.21), by fitting a simplified version of that model with only known covariates. The covariates I include are mean NO_2 concentration in each IG, as well as the two proxy measures of socio-economic deprivation, namely the percentage of people in receipt of job seekers allowance (JSA) and the natural log of the median property price (Logprice). The residuals from this model show substantial spatial autocorrelation, with significant Moran's I statistics ranging between 0.254 and 0.320 over the five years. These residuals also exhibit temporal autocorrelation, as the correlation between two consecutive period residuals are 0.659, 0.632, 0.630, 0.651 respectively (computed between the 1,207 spatial data points corresponding to 1,207 IGs for consecutive years). Therefore it is appropriate to include the random effects in Model (4.21) to allow for the spatio-temporal autocorrelation remaining in the disease counts after the covariate effects have been accounted for.

I fit four different models to the respiratory disease hospital admissions data, which differ only in the NO_2 metric included in the model. Model I and II correspond to the

FIGURE 4.6: Comparison between DEFRA and predicted NO_2 concentrations (μgm^{-3}) from Model 1A for 2010. The left panel is a scatterplot of all model-predicted versus DEFRA values while the right panel shows a spatial map of the differences between them.



spatial mean and maximum of the DEFRA concentrations, while Models III and IV relate to the spatial mean and maximum of the predicted pollution concentrations from Model 1A. The results of fitting these models are displayed in Table 4.6, which shows that $\rho \approx 0.92$ and $\gamma \approx 0.83$ indicating high spatial and temporal autocorrelation in the disease data after the covariate effects have been accounted for, validating the use of the random effects model. These results are robust to the choice of NO_2 metric used in the model. Table 4.6 also shows that the covariate effects are substantial and robust across the four models, as their 95% credible intervals do not contain the null risk value of one. This indicates that the natural log of the median property price and the percentage of people receiving job seeker allowance are significantly related to hospital admissions, with a 0.38 increase in Logprice relating to 8% lower hospital admissions while a 2.35% increase in JSA results in 20% higher hospital admission rates.

Finally, Table 4.6 displays the long-term effects of the four metrics of NO_2 on respiratory hospitalisation risk, which are presented as relative risks for a 6.84 μgm^{-3} (one standard deviation of the mean NO_2 across the 1,207 IGs) increase in concentrations. The spatial maximum of DEFRA concentrations (Model II) in each IG shows a significant

relationship with respiratory disease while the spatial mean of DEFRA concentrations (Model I) does not. Model II indicates that a 6.84 μgm^{-3} increase in maximum NO_2 exposure is associated with 2.3% higher respiratory disease hospital admissions in Scotland, whereas no relationship is observed when the spatial mean is used. This is similar to the work of Young et al. [143], who found that the risk of myocardial infarction is more highly correlated with monthly maximum ozone concentrations than the average concentrations. However, as previously discussed the DEFRA concentrations are known to be biased estimates of exposure (see Table 4.4), but the results from Models III and IV using the pollution concentrations estimated from Model 1A validate those using the DEFRA concentrations. Specifically, the spatial maximum concentrations in Model IV are associated with a significant 2.1% increased risk of disease, in comparison to a 2.3% increased disease risk using the DEFRA concentrations. Similarly, the spatial mean metric used in Models I and III show no relationship with disease risk.

TABLE 4.6: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the disease model (4.21) with four different pollution metrics. Model I and II correspond to the spatial mean and maximum of the DEFRA concentrations, while Models III and IV relate to the spatial mean and maximum of the predicted pollution concentrations from Model 1A. The regression parameters are presented as relative risks for a standard deviation increase in each covariates value, which is NO_2 6.84 μgm^{-3}, Logprice 0.38, JSA 2.35.

Parameter	**Model I**	**Model II**	**Model III**	**Model IV**
NO_2	1.009	1.023	0.993	1.021
	(0.991,1.028)	(1.008,1.038)	(0.980,1.008)	(1.004,1.037)
Logprice	0.920	0.920	0.920	0.921
	(0.910,0.931)	(0.910,0.929)	(0.909,0.929)	(0.911,0.930)
JSA	1.197	1.196	1.200	1.196
	(1.181,1.213)	(1.181,1.212)	(1.185,1.215)	(1.180,1.214)
ν^2	0.061	0.061	0.061	0.061
	(0.057,0.065)	(0.056,0.065)	(0.056,0.065)	(0.056,0.065)
ρ	0.917	0.918	0.926	0.911
	(0.877,0.951)	(0.879,0.951)	(0.891,0.956)	(0.866,0.946)
γ	0.831	0.831	0.832	0.830
	(0.795,0.867)	(0.794,0.867)	(0.797,0.867)	(0.792,0.865)

4.4.4 Sensitivity analysis

As mentioned in Section 4.3.2, the flexible spatial-temporal random effects in the health model are included to account for residual auto-correlation after accounting for the effects of covariates. These flexible random effects need to compete with the explanatory ability of the NO_2 exposure. Therefore, I test the robustness of the health associations

TABLE 4.7: Relative risk of NO_2 against various basis dimensions of the space smoothness.

Basis Dimension	**Model I**	**Model II**	**Model III**	**Model IV**
k=30	1.008	1.023	0.999	1.032
	(1.002,1.014)	(1.018,1.029)	(0.994,1.004)	(1.026,1.038)
k=40	1.009	1.025	0.996	1.027
	(1.003,1.016)	(1.019,1.031)	(0.990,1.001)	(1.021,1.034)
k=50	1.007	1.024	0.993	1.026
	(1.000,1.014)	(1.018,1.030)	(0.987,0.998)	(1.019,1.033)
k=60	1.002	1.021	0.989	1.023
	(0.995,1.009)	(1.014,1.027)	(0.983,0.994)	(1.016,1.030)

by fitting a range of generalized additive models to the data, where the random effects are replaced by smooth functions in space and time (splines) with varying levels of smoothness. Specifically, I use a linear combination of separate smooth functions for space and time, with the former being an isotropic smooth function using thin plate splines. As the data in my study contain only 5 years, the basis dimension for time can vary from 3 to 5, which actually makes little change in the smooth function and therefore is fixed at the median value 4 in the analysis. I test the robustness of the health associations against a set of different basis dimensions for the spatial smooth term, and the results are shown in Table 4.7. Table 4.7 shows that the health associations with NO_2 are robust against varying levels of control for space smoothness, as the estimates are similar to those in Table 4.6 regardless of the different levels of space smoothness.

4.5 Discussion

In this chapter, I have proposed an integrated model for estimating the long-term health effects of air pollution, that fuses DEFRA and measured pollution data to provide improved predictions of areal level pollution concentrations and hence health effects. The improvement in the pollution prediction is highlighted in Table 4.4, which shows a 25% and a 64% decrease in RMSPE compared to using a spatio-temporal random effects model and the DEFRA concentrations respectively. The epidemiological study presented in this chapter is one of the most comprehensive investigations into the effects of NO_2 concentrations on health in Scotland, as my study region is all of mainland Scotland for the five year period spanning from 2007 to 2011.

My findings show that a 6.84 μgm^{-3} increase in peak NO_2 concentrations (spatial maximum) within an IG is associated with 2.3% higher respiratory disease hospital admissions in Scotland, while no such relationship is observed with mean concentrations (spatial mean) in an IG. This suggests that the choice of spatial aggregation metric used to quantify areal level pollution concentrations has a major impact on the resulting health effect estimate, which naturally leads to the question of which metric should one use. This issue has received little attention to date in the literature, as different exposure metrics have been used in epidemiological studies (see e.g. Basu et al. [6] and Berrocal et al. [12]). However, the majority of epidemiological studies use the average (mean) concentration (see e.g. Maheswaran et al. [84]; Lee et al. [76] and Warren et al. [136]).

The second interesting finding of my research is the consistency between the estimated health effects of NO_2, when the latter is estimated using the DEFRA concentrations alone and both the measured and DEFRA concentrations. This consistency was observed when considering both the spatial mean and maximum as the aggregation functions, and suggests that the DEFRA concentrations appear reliable to use in health effect studies despite being biased.

There is a limitation of the design of the monitoring network in my study, where the monitor locations are highly clustered in the central part of the study region in Glasgow (west) and Edinburgh (east), and no monitors exist in large parts of the study region (see Figure 4.1). Therefore, the predictive performance cannot be assessed uniformly across Scotland when I evaluated the prediction performance of several exposure models using a 10-fold cross-validation. In other words, the prediction performance at rural areas where no monitors exist is unknown. However, as these areas are rural regions then NO_2 concentrations are low (away from traffic sources), so the level of uncertainty should be low and the DEFRA concentrations should be able to pick up the low background levels. This is also the key reason for using a fusion model to utilize the good spatial coverage of the DEFRA concentrations to provide estimation of pollution in these largely rural regions.

In my study the air pollution concentrations are assumed to be known and constant across each IG, by spatially aggregating predictions from my pollution model. However, these predictions are likely to contain errors and uncertainties, and in future work I will investigate these issues within a hierarchical modelling framework. A further avenue

of future work could be the investigation of the individual and joint effects of different pollutants, rather than simply considering NO_2 as was the case here.

Chapter 5

Multi-pollutant concentrations prediction

5.1 Introduction

In the previous chapter, the long-term effects of NO_2 on respiratory diseases in mainland Scotland has been investigated, in a single-pollutant health study. However, the real world is much more complex, and the air we breathe contains a complex mixture of particle and gas phase pollutants which means that we are exposed to multiple pollutants simultaneously. These pollutants might act independently or in combination (in an additive, synergistic, antagonistic, or interactive manner) to affect human health. A traditional single pollutant health study fails to account for these combined effects of pollutant mixtures. Therefore, it is necessary to extend the current single pollutant risk assessment to account for multiple pollutants. This chapter will build a new model to predict multi-pollutant concentrations simultaneously, which then can be used for the study of multi-pollutant health effects in the next chapter.

Model 4.3 in the previous chapter is a general model to deal with the prediction of a single pollutant. For predicting multiple pollutant concentrations, a possible way is to apply Model 4.3 to each pollutant separately. However, this method ignores the correlations among pollutants (e.g. pollutant dependence was shown in Shaddick and Wakefield [117], Kumar and Joseph [70] and Berrocal et al. [11], whilst in my data set the correlation between PM_{10} and NO_2 is about 0.8), with which the prediction of

one pollutant can borrow strength from the remaining pollutants. For example, Berrocal et al. [13] showed that a bivariate downscaler model, which allows for correlation between Ozone and $PM_{2.5}$, outperforms the independent downscaler model.

The aim of this chapter is to propose a multi-pollutant model for Scotland that extends Model 4.3, based on which the multi-pollutant concentrations can be predicted across mainland Scotland simultaneously and the correlation between pollutants is allowed to be used to improve the prediction. The modelling carried out in this chapter will be used to provide pollution predictions for a study investigating the health effects of multi-pollutants in the next chapter. The structure of this chapter is organized as follows. Section 5.2 introduces the data used in my study. Section 5.3 describes the multi-pollutant model then deduces the posterior distributions for the model parameters and a McMC inference scheme, and discusses issues of missing data. Section 5.4 report a simulation study for the multi-pollutant model to assess its efficacy, while Section 5.5 is a validation study comparing the performance of the multi-pollutant model with the previous single pollutant model. Section 5.6 models the real data set in my study. Section 5.7 concludes with some discussion.

5.2 Data description

As described in chapter 3, my focus remains on mainland Scotland. Note that the pollutants people breathe include NO_2, PM_{10}, O_3, $PM_{2.5}$, SO_2, CO and so on, however, the pollutants considered in this study include only NO_2 and PM_{10} due to the sparse observations for the other pollutants. For example, the number of monitoring sites for $PM_{2.5}$ from 2006 to 2010 are 0, 1, 1, 3, 5, respectively, while they are 7, 3, 2, 3, 3 for CO.

Data for both NO_2 and PM_{10}, which are the annual mean concentrations from 2006 to 2010, are from two sources, measured concentrations at a small number of locations and DEFRA data. The measured data are collected from automatic monitors (91 monitoring sites) which can be freely obtained from Air Quality in Scotland (*http://www.scottishairquality.co.uk/*). The top left of Figure 5.1 is a map of the monitoring sites for both NO_2 and PM_{10} in 2010, in which is shown that the monitors located in big cities (e.g. Glasgow and Edinburgh in the bottom central of the map) are likely

to measure both pollutants, while there are some sites away from big cities measuring NO_2 only. This is because PM_{10} mainly comes from the burning of fossil fuels in vehicles which happens mostly in big cities, while NO_2 comes mainly from not only the burning of fossil fuels but also the atmosphere by lightning and some is produced by plants, soil and water. Among the 91 monitoring sites, there are actually 29 sites having neither NO_2 nor PM_{10} observations from 2006 to 2010, which indicates that the total sites for this study is 62 rather than 91. There are in total 50 monitoring sites with NO_2 measurements for 2010 and 42 for PM_{10}, among which 33 sites have measurements for both pollutants. Similar information for the other years can be seen in Table 5.1. All these monitoring observations will be used to predict pollution concentrations across my study region.

The monitoring sites have been classified as either urban background, kerbside, roadside or rural, with simple summaries displayed in Table 5.2. As might be expected, for both PM_{10} and NO_2, the pollution levels recorded at urban locations are higher than those at rural locations, and the closer the monitoring stations are to a main road, the higher the NO_2 and PM_{10} concentrations are.

TABLE 5.1: Numbers of monitoring sites measuring NO_2 and PM_{10}.

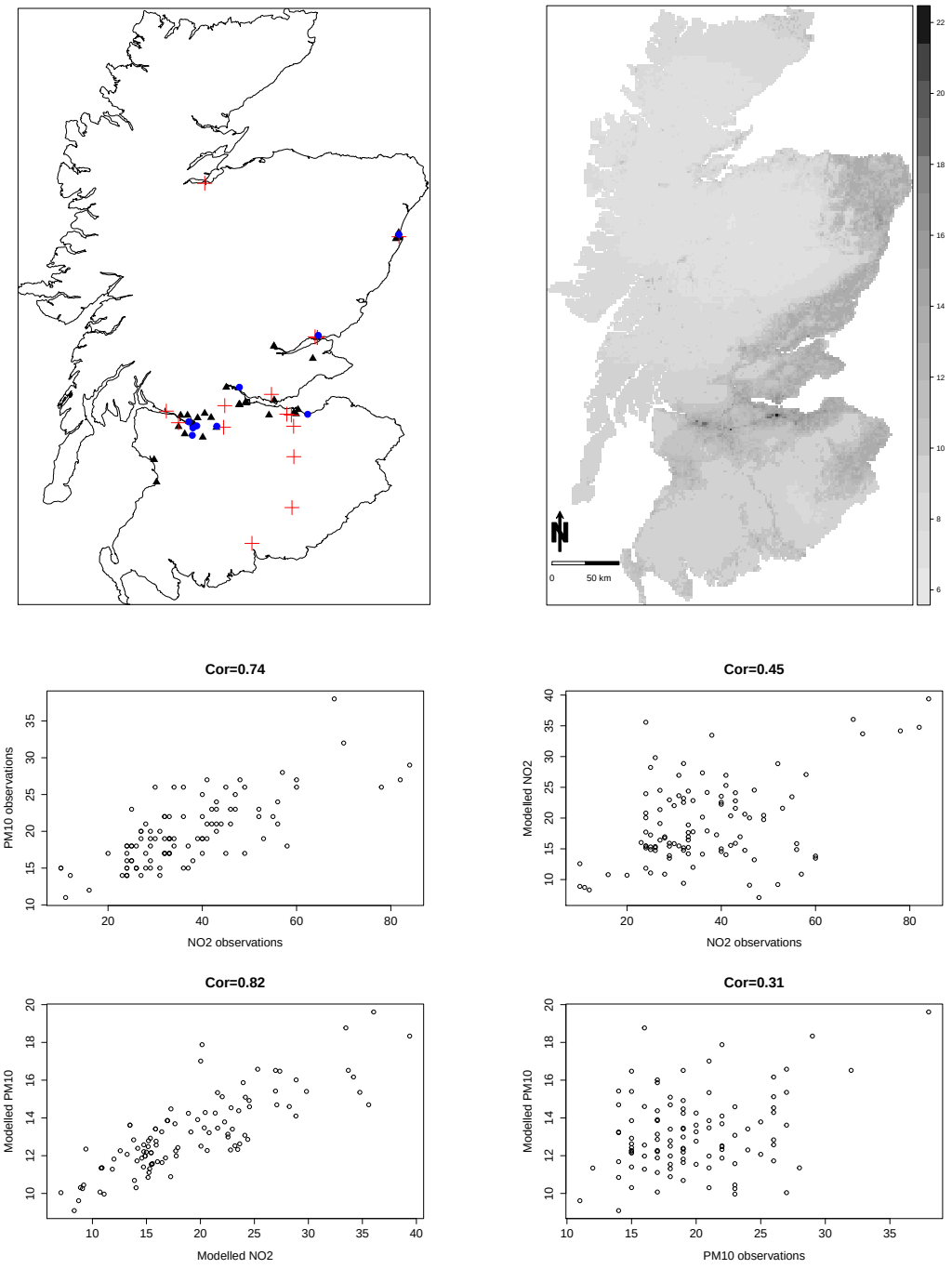
Site type	**2006**	**2007**	**2008**	**2009**	**2010**
Sites for NO_2	18	27	40	44	50
Sites for PM_{10}	12	17	29	35	42
Sites for both	10	14	22	25	33

TABLE 5.2: Summary of the monitoring data by site type and year: the numbers within the round brackets represent the number of sites in the form (NO_2, PM_{10}), while those within square brackets indicate their corresponding mean concentrations (μgm^{-3}).

Site type \ Year	**2006**	**2007**	**2008**	**2009**	**2010**
Urban Background	(3, 2)	(3, 3)	(6, 6)	(6, 6)	(6, 7)
	[27.3, 20.0]	[26.3, 17.0]	[27.0, 16.2]	[26.3, 14.1]	[26.0, 14.2]
Kerbside	(1, 1)	(4, 1)	(4, 1)	(3, 2)	(5, 2)
	[68.0, 38.0]	[64.0, 32.0]	[65.5, 27.0]	[67.3, 22.0]	[59.0, 24.0]
Roadside	(11, 8)	(15, 11)	(25, 20)	(30, 26)	(34, 32)
	[43.8, 24.1]	[42.4, 22.2]	[36.9, 20.8]	[36.2, 17.7]	[38.2, 19.2]
Rural	(3, 1)	(3, 2)	(3, 2)	(3, 1)	(3, 1)
	[8.0, 15.0]	[8.0, 10.5]	[8.3, 10.5]	[7.33, 11.0]	[9.33, 12.0]

As has been presented in Figure 5.1 the monitoring observations provide poor spatial coverage of Scotland as the major cities are well represented but the rest of the study region contains hardly any monitors. Therefore, the standard geostatistical prediction

FIGURE 5.1: Summary of the data. Top left is a map of the monitoring sites for both NO_2 and PM_{10} in 2010 (▲: common sites; red +: sites with only NO_2; blue •: sites with only PM_{10}), top right is the modelled annual average PM_{10} concentrations in 2010 (μgm^{-3}), bottom left are scatter plots between measured NO_2 and measured PM_{10}, modelled NO_2 and modelled PM_{10} for common sites, bottom right are scatter plots between measured and modelled data for common sites.



methods may not be appropriate here, due to the large distances between data locations and potential prediction locations. As a result of this poor spatial coverage, the DEFRA concentrations (also called modelled concentrations or modelled data) are also used. These data have complete spatial coverage of Scotland, but are known to have certain biases and needed to be calibrated to the measured data. These data are displayed in Figure 5.1 for PM_{10} in 2010 and show that the concentrations are much higher in the lower middle part of the country (Glasgow and Edinburgh cities), and in the east of the country where other large cities are located (e.g. Aberdeen and Dundee). The modelled concentrations for NO_2 have been shown in the previous chapter.

Note that the multi-pollutant model proposed in this chapter focuses on the use of between pollutant correlations to improve prediction, it is natural and important to investigate the correlation between NO_2 and PM_{10} in my data set. Figure 5.1 shows that the linear correlation between NO_2 and PM_{10} is strong, with 0.74 between measurements and 0.82 between modelled data (DEFRA). Figure 5.1 also shows that the linear correlation between measurements and modelled data is moderate (0.45, 0.31 for NO_2 and PM_{10}, respectively), indicating that the modelled data are a good predictor of the measurements.

Finally, temperature data and the classification of urban, background are also used in this study and they are the same as those described in chapter 4.

5.3 Methodology

In this section I propose a new Bayesian hierarchical model to predict multiple pollutant concentrations simultaneously. Rather than using a single pollutant model to do the prediction for each pollutant separately (see e.g. Vinikoor-Imler et al. [134], Vinikoor-Imler et al. [135]), the model proposed in this section allows the correlation between pollutants to improve the prediction of each pollutant by borrowing strength from the other pollutants (e.g. Berrocal et al. [11]). Unlike those methods proposed by Shaddick and Wakefield [117], Berrocal et al. [11], McMillan et al. [87] and Lawson et al. [72], which include a spatial correlation term in the model, the proposed model here does not consider the spatial correlation among the observations, as the pollution observations across mainland Scotland do not have any residual spatial correlation after accounting for

the site environment, spatially correlated modelled concentrations and the temperature data (see chapter 4).

5.3.1 Modelling

The multi-pollutant model extends the single pollutant model defined in the previous chapter. Denote the n measured pollution concentrations from monitoring sites at locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ for q pollutants for year t as $\left(\boldsymbol{X}_1^{(t)}, ..., \boldsymbol{X}_q^{(t)}\right)$, where $\boldsymbol{X}_j^{(t)} = \left(X_j^{(t)}(\boldsymbol{s}_1), \ldots, X_j^{(t)}(\boldsymbol{s}_n)\right)$, and is the set of observations over space for pollutant j in year t. The first level of the multi-pollutant model is given as,

$$\begin{bmatrix} \boldsymbol{X}_1^{(t)} \\ \boldsymbol{X}_2^{(t)} \\ ... \\ \boldsymbol{X}_q^{(t)} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{Z}_1^{(t)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & ... & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_q^{(t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1^{(t)} \\ ... \\ \boldsymbol{\beta}_q^{(t)} \end{bmatrix}, \sigma_t^2 \boldsymbol{C}_{q\times q} \otimes \boldsymbol{I}_n \right), \quad t = 1, ..., T, \quad (5.1)$$

where the measured pollution data are modelled by a linear regression model with mean $(\boldsymbol{Z}_1^{(t)}\boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{Z}_q^{(t)}\boldsymbol{\beta}_q^{(t)})$ for q pollutants for year t, respectively. Here $\left(\boldsymbol{Z}_1^{(t)}, \ldots, \boldsymbol{Z}_q^{(t)}\right)$ are the $n \times p$ design matrices, and within each are an intercept term and key covariates such as site environment indicators and the modelled data. Then $\left(\boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_q^{(t)}\right)$, each of which is a $p \times 1$ vector, are the corresponding regression parameters for year t, and are assumed to be temporally autocorrelated, following a centred multivariate first order autoregressive process. That is:

$$\begin{bmatrix} \boldsymbol{\beta}_1^{(t)} \\ ... \\ \boldsymbol{\beta}_q^{(t)} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta}_1 + \kappa(\boldsymbol{\beta}_1^{(t-1)} - \boldsymbol{\beta}_1) \\ ... \\ \boldsymbol{\beta}_q + \kappa(\boldsymbol{\beta}_q^{(t-1)} - \boldsymbol{\beta}_q) \end{bmatrix}, \tau^2 \boldsymbol{I}_{pq \times pq} \right) \quad t = 2, ..., T, \tag{5.2}$$

$$\begin{bmatrix} \boldsymbol{\beta}_1^{(1)} \\ ... \\ \boldsymbol{\beta}_q^{(1)} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta}_1 \\ ... \\ \boldsymbol{\beta}_q \end{bmatrix}, \tau^2 \boldsymbol{I}_{pq \times pq} \right),$$

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ ... \\ \boldsymbol{\beta}_q \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ ... \\ \mathbf{0} \end{bmatrix}, 1000 \boldsymbol{I}_{pq \times pq} \right).$$

The extent of this temporal dependence is captured by κ, which is assigned a uniform prior on the unit interval [0,1], $\kappa \sim \text{Uniform}[0, 1]$. If $\kappa = 0$, $(\boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_q^{(t)})$ are estimated independently for each year and are smoothed towards an overall mean value $(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q)$ for all years, while if $\kappa = 1$, $(\boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_q^{(t)})$ are temporally autocorrelated with $(\boldsymbol{\beta}_1^{(t-1)}, ..., \boldsymbol{\beta}_q^{(t-1)})$.

The covariance matrix in the likelihood model (5.1) is $\sigma_t^2 \boldsymbol{C}_{q \times q} \otimes \boldsymbol{I}_n$, which assumes constant correlations between pollutants at each monitoring site, but no such correlation across sites. $\boldsymbol{I}_n$ is an $n \times n$ identity matrix, where n is the number of monitoring sites in my study. $\boldsymbol{C}_{q \times q}$ is a covariance matrix for all pollutants at the same site, in which the element C_{ij} represents the covariance between pollutant i and j at each monitoring site. This covariance matrix is assumed to follow an Inverse-Wishart distribution, $\boldsymbol{C}_{q \times q} \sim$ Inverse-Wishart$(\nu = q, \boldsymbol{\Psi} = 100 \boldsymbol{I}_{q \times q})$ (Lawson et al. [72]). Finally, σ_t^2 is a scaling parameter to allow different levels of residual variation over time. This scaling parameter is assumed to be temporally autocorrelated via a first order random walk prior, and as it must be non-negative, the log scale is used. That is:

$$\begin{aligned} \ln(\sigma_t^2) &\sim \mathrm{N}(\ln(\sigma_{t-1}^2), \sigma^2) \quad t = 2, ..., T, \\ f(\ln(\sigma_1^2)) &\propto 1. \end{aligned} \tag{5.3}$$

Finally, I choose weakly informative conjugate prior distributions for (σ^2, τ^2) by assuming them to be Inverse-gamma distributed, $\sigma^2, \tau^2 \sim$ Inverse-Gamma$(a = 0.001, b = 0.001)$. Inference for the collection of model parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1^{(1)}, \ldots, \boldsymbol{\beta}_q^{(1)}, \ldots, \boldsymbol{\beta}_1^{(T)}, \ldots, \boldsymbol{\beta}_q^{(T)}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, \boldsymbol{C}_{q\times q}, \kappa, \sigma_1^2, \ldots, \sigma_T^2, \tau^2, \sigma^2)$ is based on McMC simulation, using both Gibbs sampling and Metropolis-Hastings steps, and I implemented the algorithm in the R programming language. Full details of the algebra for the full conditional distributions is given below.

5.3.2 Computation of the posterior distribution

The posterior distribution is the probability of the parameters given the data, which is defined as the product of the likelihood function and the prior distributions. The likelihood function for Model (5.1) is given as follows.

Denote $\boldsymbol{\Sigma}_t = \sigma_t^2 \boldsymbol{C}_{q\times q} \otimes \boldsymbol{I}_n$, $\boldsymbol{X}^{(t)} = (\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_q^{(t)})$, $\boldsymbol{Z}^{(t)} = \text{diag}(\boldsymbol{Z}_1^{(t)}, \ldots, \boldsymbol{Z}_q^{(t)})$, $\boldsymbol{\beta}^{(t)} = (\boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_q^{(t)})$, the likelihood function for Model (5.1) is given by:

$$\begin{aligned} f(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)} \mid -) &= f(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)} \mid \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, \boldsymbol{C}_{q\times q}, \kappa, \sigma_1^2, \ldots, \sigma_T^2, \tau^2, \sigma^2) \\ &= \prod_{t=1}^{T} \frac{1}{(2\pi)^{qn/2} \mid \boldsymbol{\Sigma}_t \mid^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}^{(t)} - \boldsymbol{Z}^{(t)}\boldsymbol{\beta}^{(t)})^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{X}^{(t)} - \boldsymbol{Z}^{(t)}\boldsymbol{\beta}^{(t)})\right). \end{aligned}$$

Given the prior,

$$\begin{aligned} f(\boldsymbol{\Theta}) &= f(\boldsymbol{\beta}_1^{(1)}, \ldots, \boldsymbol{\beta}_q^{(1)}, \ldots, \boldsymbol{\beta}_1^{(T)}, \ldots, \boldsymbol{\beta}_q^{(T)}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, \boldsymbol{C}_{q\times q}, \kappa, \sigma_1^2, \ldots, \sigma_T^2, \tau^2, \sigma^2) \\ &= \prod_{i=1}^{T}\prod_{j=1}^{q} f(\boldsymbol{\beta}_j^{(i)}) \prod_{j=1}^{q} f(\boldsymbol{\beta}_j) f(\boldsymbol{C}_{q\times q}) f(\kappa) \prod_{i=1}^{T} f(\sigma_i^2) f(\tau^2) f(\sigma^2) \\ &= \prod_{j=1}^{q} \text{N}(\boldsymbol{\beta}_j^{(1)} \mid \boldsymbol{\beta}_j, \tau^2 \boldsymbol{I}_p) \prod_{i=2}^{T}\prod_{j=1}^{q} \text{N}(\boldsymbol{\beta}_j^{(i)} \mid \boldsymbol{\beta}_j + \kappa(\boldsymbol{\beta}_j^{(i-1)} - \boldsymbol{\beta}_j), \tau^2 \boldsymbol{I}_p) \prod_{j=1}^{q} \text{N}(\boldsymbol{\beta}_j \mid \boldsymbol{0}, 1000\boldsymbol{I}_p) \text{IW}(\boldsymbol{C}_{q\times q} \mid \nu, \boldsymbol{\Psi}) \text{U}(\kappa \mid 0, 1) \\ &\times \prod_{i=2}^{T} \text{N}(\ln(\sigma_i^2) \mid \ln(\sigma_{i-1}^2), \sigma^2) \text{IG}(\tau^2 \mid a, b) \text{IG}(\sigma^2 \mid a, b) \end{aligned}$$

The parameters are updated by using the full conditional distributions as follows (denote η_{ij} as the ijth element of $\boldsymbol{C}_{q\times q}^{-1}$).

$f(\boldsymbol{C}_{q\times q} \mid -) \propto \text{Inverse-Wishart}\left(\nu + nt, \boldsymbol{\Psi} + \Sigma_{t=1}^{T} \frac{\boldsymbol{M}_{q\times q}^{(t)}}{\sigma_t^2}\right)$, where $\boldsymbol{M}_{ij}^{(t)} = \Sigma\left(\text{diag}\left((\boldsymbol{X}_i^{(t)} - \boldsymbol{Z}_i^{(t)}\boldsymbol{\beta}_i^{(t)})(\boldsymbol{X}_j^{(t)} - \boldsymbol{Z}_j^{(t)}\boldsymbol{\beta}_j^{(t)})^\top\right)\right)$,

$$f(\tau^2 \mid -) \propto \mathrm{IG}(a + \tfrac{qpt}{2}, b + \frac{\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}(\boldsymbol{\beta}_i^{(1)}-\boldsymbol{\beta}_i)^\top(\boldsymbol{\beta}_j^{(1)}-\boldsymbol{\beta}_j)}{2} + \frac{\Sigma_{t=2}^{T}\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}\left(\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i-\kappa(\boldsymbol{\beta}_i^{(t-1)}-\boldsymbol{\beta}_i)\right)^\top\left(\boldsymbol{\beta}_j^{(t)}-\boldsymbol{\beta}_j-\kappa(\boldsymbol{\beta}_j^{(t-1)}-\boldsymbol{\beta}_j)\right)}{2}),$$

$$f(\sigma^2 \mid -) \propto \mathrm{IG}\left(a + \tfrac{1}{2}(T-1), b + \tfrac{1}{2}\Sigma_{t=2}^{T}\left(\ln(\sigma_t^2) - \ln(\sigma_{t-1}^2)\right)^2\right),$$

$$\ln f(\sigma_1^2 \mid -) \propto -\tfrac{qn}{2}\ln(\sigma_1^2) - \tfrac{1}{2\sigma^2}[\ln(\sigma_2^2) - \ln(\sigma_1^2)]^2 - \tfrac{1}{2\sigma_1^2}\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}\eta_{ij}\left(\boldsymbol{X}_i^{(1)} - \boldsymbol{Z}_i^{(1)}\boldsymbol{\beta}_i^{(1)}\right)^\top\left(\boldsymbol{X}_j^{(1)} - \boldsymbol{Z}_j^{(1)}\boldsymbol{\beta}_j^{(1)}\right),$$

$$\ln f(\sigma_T^2 \mid -) \propto -\tfrac{qn}{2}\ln(\sigma_T^2) - \tfrac{1}{2\sigma^2}[\ln(\sigma_T^2) - \ln(\sigma_{T-1}^2)]^2 - \tfrac{1}{2\sigma_T^2}\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}\eta_{ij}\left(\boldsymbol{X}_i^{(T)} - \boldsymbol{Z}_i^{(T)}\boldsymbol{\beta}_i^{(T)}\right)^\top\left(\boldsymbol{X}_j^{(T)} - \boldsymbol{Z}_j^{(T)}\boldsymbol{\beta}_j^{(T)}\right),$$

$$\ln f(\sigma_t^2 \mid -) \propto -\tfrac{qn}{2}\ln(\sigma_t^2) - \tfrac{1}{2\sigma^2}[\ln(\sigma_t^2) - \ln(\sigma_{t-1}^2)]^2 - \tfrac{1}{2\sigma^2}[\ln(\sigma_{t+1}^2) - \ln(\sigma_t^2)]^2 - \tfrac{1}{2\sigma_t^2}\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}\eta_{ij}\left(\boldsymbol{X}_i^{(t)} - \boldsymbol{Z}_i^{(t)}\boldsymbol{\beta}_i^{(t)}\right)^\top\left(\boldsymbol{X}_j^{(t)} - \boldsymbol{Z}_j^{(t)}\boldsymbol{\beta}_j^{(t)}\right), 1 < t < T,$$

$$f(\boldsymbol{\beta}_i \mid -) \propto \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$
$$\text{where } \boldsymbol{\mu}_i = \left[\left(\tfrac{1}{\tau^2} + \tfrac{1}{1000} + \tfrac{(1-\kappa)^2(T-1)}{\tau^2}\right)\boldsymbol{I}_p\right]^{-1}\left[\frac{\boldsymbol{\beta}_i}{\tau^2} + \frac{\Sigma_{j=1}^{q}(\boldsymbol{\beta}_j^{(1)}-\boldsymbol{\beta}_j)}{\tau^2} + \frac{\boldsymbol{\beta}_i}{1000} - \frac{\Sigma_{j=1}^{q}\boldsymbol{\beta}_j}{1000} + \frac{(T-1)(1-\kappa)^2\boldsymbol{\beta}_i}{\tau^2} + \Sigma_{t=2}^{T}\frac{\Sigma_{j=1}^{q}(1-\kappa)(\boldsymbol{\beta}_j^{(t)}-\kappa\boldsymbol{\beta}_j^{(t-1)}-(1-\kappa)\boldsymbol{\beta}_j)}{\tau^2}\right]$$
$$\boldsymbol{\Sigma}_i = \left[\left(\tfrac{1}{\tau^2} + \tfrac{1}{1000} + \tfrac{(1-\kappa)^2(T-1)}{\tau^2}\right)\boldsymbol{I}_p\right]^{-1},$$

$$f(\boldsymbol{\beta}_i^{(T)} \mid -) \propto \mathrm{N}(\boldsymbol{\mu}_i^{(T)}, \boldsymbol{\Sigma}_i^{(T)}),$$
$$\text{where } \boldsymbol{\mu}_i^{(T)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(T)})^\top\boldsymbol{Z}_i^{(T)}}{\sigma_T^2} + \frac{\boldsymbol{I}_p}{\tau^2}\right]^{-1}\left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(T)})^\top\boldsymbol{Z}_i^{(T)}\boldsymbol{\beta}_i^{(T)}}{\sigma_T^2} + \frac{\Sigma_{j=1}^{q}\eta_{ij}((\boldsymbol{Z}_i^{(T)})^\top\boldsymbol{X}_j^{(T)} - (\boldsymbol{Z}_i^{(T)})^\top\boldsymbol{Z}_j^{(T)}\boldsymbol{\beta}_j^{(T)})}{\sigma_T^2} + \frac{\boldsymbol{\beta}_i^{(T)}}{\tau^2} + \frac{\Sigma_{j=1}^{q}((1-\kappa)\boldsymbol{\beta}_j + \kappa\boldsymbol{\beta}_j^{(T-1)} - \boldsymbol{\beta}_j^{(T)})}{\tau^2}\right]$$
$$\boldsymbol{\Sigma}_i^{(T)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(T)})^\top\boldsymbol{Z}_i^{(T)}}{\sigma_T^2} + \frac{\boldsymbol{I}_p}{\tau^2}\right]^{-1},$$

$$f(\boldsymbol{\beta}_i^{(1)} \mid -) \propto \mathrm{N}(\boldsymbol{\mu}_i^{(1)}, \boldsymbol{\Sigma}_i^{(1)}),$$
$$\text{where } \boldsymbol{\mu}_i^{(1)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(1)})^\top\boldsymbol{Z}_i^{(1)}}{\sigma_1^2} + \frac{\boldsymbol{I}_p}{\tau^2} + \frac{\kappa^2\boldsymbol{I}_p}{\tau^2}\right]^{-1}\left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(1)})^\top\boldsymbol{Z}_i^{(1)}\boldsymbol{\beta}_i^{(1)}}{\sigma_1^2} + \frac{\Sigma_{j=1}^{q}\eta_{ij}\left((\boldsymbol{Z}_i^{(1)})^\top\boldsymbol{X}_j^{(1)} - (\boldsymbol{Z}_i^{(1)})^\top\boldsymbol{Z}_j^{(1)}\boldsymbol{\beta}_j^{(1)}\right)}{\sigma_1^2} + \frac{\boldsymbol{\beta}_i^{(1)}}{\tau^2} + \frac{\Sigma_{j=1}^{q}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{(1)})}{\tau^2} + \frac{\kappa^2\boldsymbol{\beta}_i^{(1)}}{\tau^2} + \frac{\Sigma_{j=1}^{q}(\kappa\boldsymbol{\beta}_j^{(2)} - \kappa(1-\kappa)\boldsymbol{\beta}_j - \kappa^2\boldsymbol{\beta}_j^{(1)})}{\tau^2}\right]$$
$$\boldsymbol{\Sigma}_i^{(1)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(1)})^\top\boldsymbol{Z}_i^{(1)}}{\sigma_1^2} + \frac{\boldsymbol{I}_p}{\tau^2} + \frac{\kappa^2\boldsymbol{I}_p}{\tau^2}\right]^{-1},$$

$f(\boldsymbol{\beta}_i^{(t)} \mid -) \propto \mathrm{N}(\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}), 1 < t < T,$

where $\boldsymbol{\mu}_i^{(t)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(t)})^\top \boldsymbol{Z}_i^{(t)}}{\sigma_t^2} + \frac{\boldsymbol{I}_p}{\tau^2} + \frac{\kappa^2 \boldsymbol{I}_p}{\tau^2}\right]^{-1} \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(t)})^\top \boldsymbol{Z}_i^{(t)} \boldsymbol{\beta}_i^{(t)}}{\sigma_t^2} + \frac{\Sigma_{j=1}^q \eta_{ij}\left((\boldsymbol{Z}_i^{(t)})^\top \boldsymbol{X}_j^{(t)} - (\boldsymbol{Z}_i^{(t)})^\top \boldsymbol{Z}_j^{(t)} \boldsymbol{\beta}_j^{(t)}\right)}{\sigma_t^2} + \frac{\boldsymbol{\beta}_i^{(t)}}{\tau^2} + \frac{\Sigma_{j=1}^q \left((1-\kappa)\boldsymbol{\beta}_j + \kappa \boldsymbol{\beta}_j^{(t-1)} - \boldsymbol{\beta}_j^{(t)}\right)}{\tau^2} + \frac{\kappa^2 \boldsymbol{\beta}_i^{(t)}}{\tau^2} + \frac{\Sigma_{j=1}^q \left(\kappa \boldsymbol{\beta}_j^{(t+1)} - \kappa(1-\kappa)\boldsymbol{\beta}_j - \kappa^2 \boldsymbol{\beta}_j^{(t)}\right)}{\tau^2}\right]$

$\boldsymbol{\Sigma}_i^{(t)} = \left[\frac{\eta_{ii}(\boldsymbol{Z}_i^{(t)})^\top \boldsymbol{Z}_i^{(t)}}{\sigma_t^2} + \frac{\boldsymbol{I}_p}{\tau^2} + \frac{\kappa^2 \boldsymbol{I}_p}{\tau^2}\right]^{-1}$

$f(\kappa \mid -) \propto \mathrm{N}(\mu_\kappa, \sigma_\kappa)$, where $\mu_\kappa = \frac{\Sigma_{t=2}^T \Sigma_{i=1}^q \Sigma_{j=1}^q \left((\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i)^\top (\boldsymbol{\beta}_j^{(t-1)} - \boldsymbol{\beta}_j) + (\boldsymbol{\beta}_i^{(t-1)} - \boldsymbol{\beta}_i)^\top (\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_j)\right)}{2\Sigma_{t=2}^T \Sigma_{i=1}^q \Sigma_{j=1}^q (\boldsymbol{\beta}_i^{(t-1)} - \boldsymbol{\beta}_i)^\top (\boldsymbol{\beta}_j^{(t-1)} - \boldsymbol{\beta}_j)}$, $\sigma_\kappa = \frac{\tau^2}{\Sigma_{t=2}^T \Sigma_{i=1}^q \Sigma_{j=1}^q (\boldsymbol{\beta}_i^{(t-1)} - \boldsymbol{\beta}_i)^\top (\boldsymbol{\beta}_j^{(t-1)} - \boldsymbol{\beta}_j)}$

5.3.3 Dealing with missing data

As there are in total 62 monitoring sites used to measure NO_2 and PM_{10} during 2006-2010, it is expected to have in total 620 (2 pollutants $\times$ 62 sites $\times$ 5 years) observations. However, according to the map of the monitoring sites in Figure 5.1 and the summaries of monitoring sites in table 5.1, some sites measure NO_2 or PM_{10} only and the number of monitors for each year is also different. Table 5.1 shows that there are actually only 314 observations (sum of the first two rows) for both pollutants during 2006-2010. The proportion of missing data is about 49%.

There are a few reasons that could lead to these missing data. Firstly, the urban air pollution monitoring sites are usually used to detect noncompliance with air quality standards [EPA (2006)], so they are preferentially located according to different pollutants. Secondly, the monitoring sites for a specified pollutant are also changed over time, with some new sites being added to the network while some of the existing sites are removed.

In my study, the model proposed in this chapter is a Bayesian model and a well known method to deal with missing data under a Bayesian framework is to treat the missing data as parameters, then update them using McMC (Tan et al. [124]). This goal can be achieved based on the multivariate normal distribution theory. Specifically, denote a vector of all the measurements for one year as $\boldsymbol{Y}_2$, all the missing monitoring data for the same year as $\boldsymbol{Y}_1$, then we can have,

$$\begin{aligned} \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} &\sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) \\ \boldsymbol{Y}_1 \mid \boldsymbol{Y}_2 &\sim \mathrm{N}\left(\boldsymbol{\theta}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}[\boldsymbol{Y}_2 - \boldsymbol{\theta}_2], \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right) \end{aligned} \tag{5.4}$$

As it is computationally expensive to compute the inverse matrix for $\boldsymbol{\Sigma}_{22}$ in each iteration of the McMC scheme, and it is known from the previous study that the spatial correlation between the monitoring data are very weak, therefore an approximate method used to update the missing observation at each site is adopted. This approximate method assumes the distribution of a missing value at a monitoring site only depends on the other pollutants at the same site rather than all the observations across all sites. For

example, for a monitoring site k with NO_2 measurement but not PM_{10} (the case of two pollutants), PM_{10} measurement is missing at site k and its prediction only depends on the NO_2 measurement at site k.

5.4 Simulation study

In this section, I present a simulation study to evaluate the performance of the multi-pollutant model proposed in the previous section. This simulation contains three parts, with the first part describing how the simulated data are generated. The second part describes the simulation method and the last part presents the results of the simulation.

5.4.1 Data generation

Simulated pollution data are generated from Model (5.1) for two pollutants, NO_2 ($\boldsymbol{X}_1^{(t)}$) and PM_{10} ($\boldsymbol{X}_2^{(t)}$). As there are more common monitoring sites (measuring both NO_2 and PM_{10}) in 2010 than the other years (see Table 5.1), I use the covariate data corresponding to these monitoring sites to form the design matrices ($\boldsymbol{Z}_1^{(t)}, \boldsymbol{Z}_2^{(t)}$) in Model (5.1) in this simulation study. Each design matrix consists of an intercept term and variables including indicator variables for kerbside, roadside and rural site type, modelled data and temperature. The regression parameters for each year, $\left(\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}\right)$, were generated from a temporal model (5.2) based on the initial value $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where the temporal correlation control parameter κ is fixed at 0.6 which is similar to the estimate from the single pollutant Model (4.3) applied to the NO_2 data in the last chapter. τ^2 which represents the white noise of the regression parameters over time is fixed at 0.1 and the overall mean values $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ are also fixed and shown in Table 5.3. In the table, β_{11} refers to the first element of $\boldsymbol{\beta}_1$ while β_{21} is the first element of $\boldsymbol{\beta}_2$. The values of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ are the maximum likelihood estimates obtained by fitting a normal linear model to NO_2 and PM_{10} data for 2010 (both measured and modelled pollution data are on the log scale).

Furthermore, in Model (5.1) σ_t^2 was also generated from a temporal Model (5.3) on the log scale based on an initial value $\sigma_1^2 = 0.06$, with the white noise $\sigma^2 = 0.01$, and $\boldsymbol{C}_{2\times 2} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$, which indicates the variances of both NO_2 and PM_{10} for the first

period are $\sigma_1^2 \times C_{11} = \sigma_1^2 \times C_{22} = 0.06$, and the correlation between NO_2 and PM_{10} is 0.7. The equal variances assumption of NO_2 and PM_{10} is a simplification of the real data set in my study, as in 2010, the variances of $\log(NO_2)$ and $\log(PM_{10})$ are 0.102 and 0.040, respectively. The correlation value is similar to the correlation between NO_2 and PM_{10} from the real data set in 2010 which is 0.74.

TABLE 5.3: Simulation settings of regression parameters for Model (5.1)

Covariate	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
Intercept	β_{11}= 3.00	β_{21}=2.14
Kerbside	β_{12}=0.67	β_{22}=0.48
Roadside	β_{13}=0.33	β_{23}=0.28
Rural	β_{14}=-0.23	β_{24}=-0.08
Modelled data	β_{15}=0.41	β_{25}=0.39
Temperature	β_{16}=-0.12	β_{26}=-0.06

5.4.2 Simulation method

I use the settings above to generate 100 simulated data sets, comprising $\left(\boldsymbol{X}_1^{(t)}, \boldsymbol{X}_2^{(t)}\right)$ for $t = 1, ..., 5$. Each simulated data set is then used to fit the multi-pollutant model with 50,000 iterations (with 20,000 as burn-in iterations, after which the chain is checked for convergence). Model inference is obtained based on 30,000 posterior samples. In order to assess the performance of the multi-pollutant model, three statistics are calculated for each model parameter, namely: bias, root mean square error (RMSE) and the coverage probabilities of the 95% credible interval (CI). Note that in this simulation study, some parameters are fixed in each of the simulated data sets as mentioned above (e.g. $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$), however, some vary in each simulation (e.g. $\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}$). The latter are allowed to vary as they are generated from a specific distribution in each generation of the simulated data. Therefore, the bias and RMSE in this simulation study are given by

$$\begin{aligned} \text{Bias}(\hat{\theta}) &= \frac{1}{m}\Sigma_{i=1}^{m}(\hat{\theta}_i - \theta_i) \qquad m = 100 \\ \text{RMSE}(\hat{\theta}) &= \sqrt{\frac{1}{m}\Sigma_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2} \qquad m = 100, \end{aligned}$$

where θ_i is the true value of any parameter in the ith simulated data set and $\hat{\theta}_i$ is its estimate. Finally, the CI coverage is the percentage of the 95% CI containing the true parameter value.

5.4.3 Simulation results

The simulation results are shown in Table 5.4, 5.5 and 5.6.

TABLE 5.4: The bias of each parameter from the simulation study of Model (5.1).

Fixed parameters	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	0.04	-0.04	-0.01	-0.01	-0.03	-0.01	0.02	-0.05	-0.02	-0.00	-0.03	-0.00
Fixed parameters	σ^2	κ	τ^2									
	0.02	-0.06	-0.01									
Varied parameters	$\beta_{11}^{(t)}$	$\beta_{12}^{(t)}$	$\beta_{13}^{(t)}$	$\beta_{14}^{(t)}$	$\beta_{15}^{(t)}$	$\beta_{16}^{(t)}$	$\beta_{21}^{(t)}$	$\beta_{22}^{(t)}$	$\beta_{23}^{(t)}$	$\beta_{24}^{(t)}$	$\beta_{25}^{(t)}$	$\beta_{26}^{(t)}$
t=1	0.06	-0.03	0.00	-0.04	-0.00	-0.01	0.06	-0.03	0.00	-0.04	-0.00	-0.01
t=2	0.06	-0.04	-0.03	-0.02	-0.01	-0.00	0.06	-0.04	-0.03	-0.02	-0.01	-0.00
t=3	0.06	-0.00	-0.02	-0.01	0.02	-0.01	0.06	-0.00	-0.02	-0.01	0.02	-0.01
t=4	0.04	-0.02	-0.02	-0.04	-0.02	0.00	0.04	-0.02	-0.02	-0.04	-0.02	0.00
t=5	0.06	-0.01	-0.02	-0.06	-0.02	0.00	0.06	-0.01	-0.02	-0.06	-0.02	0.00
Covariance	$\sigma_1^2 C_{ij}$	$\sigma_2^2 C_{ij}$	$\sigma_3^2 C_{ij}$	$\sigma_4^2 C_{ij}$	$\sigma_5^2 C_{ij}$							
i=j=1	0.002	0.001	0.001	-0.000	-0.001							
i=1,j=2	0.001	0.001	0.000	-0.000	-0.001							
i=j=2	0.003	0.000	0.001	0.001	-0.000							

TABLE 5.5: The RMSE of each parameter from the simulation study of Model (5.1).

Fixed parameters	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	0.50	0.29	0.26	0.26	0.27	0.25	0.61	0.25	0.26	0.26	0.26	0.24
Fixed parameters	σ^2	κ	τ^2									
	0.04	0.20	0.03									
Varied parameters	$\beta_{11}^{(t)}$	$\beta_{12}^{(t)}$	$\beta_{13}^{(t)}$	$\beta_{14}^{(t)}$	$\beta_{15}^{(t)}$	$\beta_{16}^{(t)}$	$\beta_{21}^{(t)}$	$\beta_{22}^{(t)}$	$\beta_{23}^{(t)}$	$\beta_{24}^{(t)}$	$\beta_{25}^{(t)}$	$\beta_{26}^{(t)}$
t=1	0.51	0.18	0.13	0.20	0.09	0.05	0.51	0.18	0.13	0.20	0.09	0.05
t=2	0.49	0.15	0.11	0.19	0.11	0.05	0.49	0.15	0.11	0.19	0.11	0.05
t=3	0.48	0.16	0.12	0.20	0.09	0.05	0.48	0.16	0.12	0.20	0.09	0.05
t=4	0.45	0.17	0.13	0.20	0.10	0.05	0.45	0.17	0.13	0.20	0.10	0.05
t=5	0.45	0.17	0.12	0.26	0.09	0.06	0.45	0.17	0.12	0.26	0.09	0.06
Covariance	$\sigma_1^2 C_{ij}$	$\sigma_2^2 C_{ij}$	$\sigma_3^2 C_{ij}$	$\sigma_4^2 C_{ij}$	$\sigma_5^2 C_{ij}$							
i=j=1	0.010	0.008	0.009	0.008	0.010							
i=1,j=2	0.008	0.007	0.007	0.007	0.008							
i=j=2	0.010	0.009	0.009	0.008	0.009							

TABLE 5.6: The CI coverage (%) for each parameter from the simulation study of Model (5.1).

Fixed parameters	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	94	94	95	98	91	93	95	94	94	98	97	97
Fixed parameters	σ^2	κ	τ^2									
	97	100	98									
Varied parameters	$\beta_{11}^{(t)}$	$\beta_{12}^{(t)}$	$\beta_{13}^{(t)}$	$\beta_{14}^{(t)}$	$\beta_{15}^{(t)}$	$\beta_{16}^{(t)}$	$\beta_{21}^{(t)}$	$\beta_{22}^{(t)}$	$\beta_{23}^{(t)}$	$\beta_{24}^{(t)}$	$\beta_{25}^{(t)}$	$\beta_{26}^{(t)}$
t=1	94	93	89	97	98	94	94	93	89	97	98	94
t=2	97	97	96	98	97	93	97	97	96	98	97	93
t=3	95	94	96	98	98	93	95	94	96	98	98	93
t=4	96	94	94	94	98	94	96	94	94	94	98	94
t=5	96	95	92	93	98	96	96	95	92	93	98	96
Covariance	$\sigma_1^2 C_{ij}$	$\sigma_2^2 C_{ij}$	$\sigma_3^2 C_{ij}$	$\sigma_4^2 C_{ij}$	$\sigma_5^2 C_{ij}$							
i=j=1	98	98	99	98	98							
i=1,j=2	95	98	98	99	98							
i=j=2	96	97	98	99	100							

TABLE 5.7: Bias, RMSE and 95% coverage (the same order in each bracket) for the covariance in the simulation study of Model (5.1) for more levels of correlation between NO_2 and PM_{10}.

Correlation	**Index**	$\sigma_1^2 C_{ij}$	$\sigma_2^2 C_{ij}$	$\sigma_3^2 C_{ij}$	$\sigma_4^2 C_{ij}$	$\sigma_5^2 C_{ij}$
Corr=0.1	i=j=1	(0.001, 0.009, 98)	(0.000, 0.009, 94)	(0.000, 0.009, 92)	(0.001, 0.009, 97)	(0.000, 0.011, 95)
	i=1,j=2	(-0.000, 0.005, 98)	(0.000, 0.005, 95)	(0.000, 0.005, 95)	(0.000, 0.005, 94)	(0.000, 0.005, 97)
	i=j=2	(0.002, 0.009, 100)	(0.001, 0.008, 97)	(0.001, 0.009, 97)	(0.002, 0.008, 97)	(0.001, 0.011, 93)
Corr=0.5	i=j=1	(0.000, 0.009, 98)	(-0.000, 0.008, 100)	(-0.001, 0.009, 97)	(-0.001, 0.008, 99)	(-0.001, 0.010, 98)
	i=1,j=2	(-0.001, 0.007, 95)	(-0.001, 0.006, 95)	(-0.002, 0.006, 95)	(-0.001, 0.006, 95)	(-0.002, 0.007, 97)
	i=j=2	(0.001, 0.009, 96)	(0.000, 0.009, 98)	(-0.000, 0.009, 94)	(-0.000, 0.009, 96)	(-0.002, 0.010, 97)
Corr=0.9	i=j=1	(0.001, 0.010, 94)	(0.000, 0.009, 93)	(-0.000, 0.009, 94)	(0.000, 0.009, 97)	(-0.000, 0.009, 98)
	i=1,j=2	(0.001, 0.010, 95)	(0.000, 0.009, 95)	(-0.000, 0.008, 95)	(0.000, 0.009, 94)	(-0.000, 0.009, 98)
	i=j=2	(0.002, 0.011, 93)	(0.001, 0.009, 94)	(0.000, 0.009, 96)	(0.001, 0.009, 95)	(0.000, 0.009, 97)

TABLE 5.8: Comparison of bias, RMSE and 95% coverage (the same order in each bracket) for fixed coefficients between the simulation study of Model (5.1) (round brackets) and single-pollutant model (4.3) (square brackets) for more levels of correlation between NO_2 and PM_{10}.

Parameter	Corr=0.1	Corr=0.5	Corr=0.9
β_{11}	(0.04, 0.53, 93) [2.22, 2.53, 100]	(-0.05, 0.55, 95) [2.36, 2.70, 100]	(0.06, 0.47, 97) [2.35, 2.68, 100]
β_{12}	(-0.05, 0.25, 94) [-0.03, 1.11, 70]	(0.01, 0.25, 96) [0.02, 1.09, 70]	(-0.02, 0.27, 93) [-0.06, 1.13, 74]
β_{13}	(0.01, 0.26, 94) [-0.32, 1.14, 78]	(0.00, 0.28, 92) [-0.32, 1.14, 80]	(-0.04, 0.26, 97) [-0.35, 1.17, 80]
β_{14}	(0.05, 0.34, 89) [-0.90, 1.43, 76]	(-0.02, 0.31, 93) [-0.92, 1.46, 71]	(0.04, 0.31, 93) [-0.94, 1.49, 72]
β_{15}	(0.00, 0.27, 92) [-0.28, 1.20, 95]	(-0.03, 0.25, 95) [-0.31, 1.13, 96]	(-0.02, 0.22, 94) [-0.32, 1.16, 92]
β_{16}	(0.33, 0.25, 97) [-0.78, 1.35, 83]	(-0.01, 0.25, 91) [-0.81, 1.36, 82]	(0.04, 0.25, 94) [-0.81, 1.33, 85]
β_{21}	(-0.01, 0.72, 92) [1.62, 1.91, 100]	(-0.15, 0.56, 99) [1.67, 1.97, 100]	(0.03, 0.57, 94) [1.76, 2.01, 100]
β_{22}	(0.03, 0.28, 95) [-0.09, 0.81, 82]	(0.00, 0.28, 95) [-0.06, 0.84, 80]	(-0.06, 0.31, 90) [0.00, 0.81, 81]
β_{23}	(-0.03, 0.25, 98) [-0.21, 0.84, 80]	(-0.00, 0.26, 94) [-0.21, 0.81, 84]	(0.00, 0.27, 95) [-0.21, 0.80, 85]
β_{24}	(0.05, 0.31, 91) [-0.60, 1.02, 73]	(0.01, 0.32, 90) [-0.60, 1.03, 77]	(-0.03, 0.28, 98) [-0.62, 0.98, 80]
β_{25}	(0.05, 0.29, 96) [-0.17, 0.90, 100]	(0.03, 0.33, 92) [-0.15, 0.80, 100]	(-0.05, 0.30, 92) [-0.17, 0.84, 100]
β_{26}	(-0.03, 0.26, 90) [-0.56, 0.94, 90]	(-0.03, 0.28, 91) [-0.58, 0.95, 89]	(0.00, 0.24, 97) [-0.57, 0.97, 89]

Table 5.4 shows that the biases for all the regression parameters are close to zero so the model parameter estimates are unbiased. Note that the bottom left of Table 5.4 refers to the bias of $\sigma_t^2 C_{ij}$ (covariance matrix element) rather than for σ_t^2 and C_{ij} separately, because $\boldsymbol{C}_{q\times q}$ is updated by an Inverse-Wishart distribution which produces only a covariance matrix rather than a correlation matrix, and this covariance matrix is possibly on a different scale compared to the original values because the scaling parameter σ_t^2 can adjust this difference. All the estimated elements of the covariance matrix have ignoreable biases, indicating the proposed multi-pollutant model can inform both the variance of each pollutant and their correlation based on the simulated data. For unbiased estimators, the RMSE measures the amount of variation in the estimate around the true value, with smaller values indicating more precise estimation. Table 5.5 displays the RMSE for each model parameter in the simulation study of the multi-pollutant model, which shows that the intercepts in the model $(\beta_{11}, \beta_{21}, \beta_{11}^{(t)}, \beta_{21}^{(t)})$ have much higher uncertainty than the other regression parameters. This is expected as in a simple linear model, the intercept is sensitive to the change of slope of the fitted line. The RMSE of the elements of the covariance matrix are very low (ranging from 0.007 to 0.010) indicating that the variance of each pollutant and their correlation are estimated precisely. In addition, the table also shows that the multi-pollutant model has low RMSE values for the white noise σ^2 and τ^2.

Table 5.6 presents the coverage for each parameter, which shows that all the parameters are estimated well because the coverages are quite close to their nominal 95% level. Note that the CI coverage of $\beta_{13}^{(1)}$ and $\beta_{23}^{(1)}$ are a little lower than those from the other years, which is likely because the variance of their posteriors are smaller compared to those from the other four years. For example, the variance of $\beta_{13}^{(1)}$ is 0.098 while they are 0.134, 0.146, 0.169, 0.133 for $\beta_{13}^{(2)}$, $\beta_{13}^{(3)}$, $\beta_{13}^{(4)}$, $\beta_{13}^{(5)}$, respectively. On the other hand, the CI coverage for $\sigma_5^2 C_{22}$ is a little higher than those from the other years, which is likely because the variance of its posterior samples is higher than those from the other years (the variance of $\sigma_5^2 C_{22}$ is 0.00022 while they are 0.00014, 0.00012, 0.00009, 0.00010 for $\sigma_1^2 C_{22}$, $\sigma_2^2 C_{22}$, $\sigma_3^2 C_{22}$, $\sigma_4^2 C_{22}$, respectively). The high CI coverage of κ is also likely due to the high variance of the posterior samples.

In order to test the performance of the multi-pollutant model (5.1), I consider the extra simulations with different levels of correlation (Corr=0.1, 0.5, 0.9) between NO_2 and PM_{10} and also compare their results with those from the single-pollutant model (4.3).

Table 5.7 shows that the bias, RMSE and coverage for the covariance for different levels of correlation between NO_2 and PM_{10} are similar to those with Corr=0.7. Therefore, the multi-pollutant model works for different levels of correlation. Table 5.8 compares the performance between the multi-pollutant model and single pollutant model, which suggests that the former outperforms the latter in terms of bias, RMSE and coverage.

5.5 Validation study

In this section, a validation study is presented, that compares the performance of the multi-pollutant model (5.1) and the single pollutant model (4.3). As the multi-pollutant model proposed in this chapter focuses on using between pollutant correlations to improve prediction, I use the observations from the common sites as the test data of this validation study. This enables the information regarding one pollutant to help improve the prediction of the other pollutant. For computational efficiency, only the common sites for 2010 have been used. Specifically, I run a leave one out cross validation for those common monitoring sites excluding 2 kerbside sites and 1 rural site. That is, for each common site in 2010 I leave out the PM_{10} concentration only, and use all the remaining observations for both NO_2 and PM_{10} from 2006-2010 to predict it. This process is then repeated for NO_2. In running the single pollutant model, only the observations at each site in 2010 are left out. The results of the validation study are shown in Table 5.9 which indicates that the multi-pollutant model outperforms the single pollutant model.

TABLE 5.9: Bias, root mean square prediction error and coverage probabilities from a leave one out cross validation excercise for the single pollutant model (4.3) and the multipollutant model (5.1), based on the the common sites in 2010.

Model	**Bias**	**RMSPE**	**CI Coverage**
Single pollutant for NO_2	-0.008	0.248	96.6
Multi-pollutant for NO_2	-0.006	0.213	96.6
Single pollutant for PM_{10}	-0.015	0.160	90.0
Multi-pollutant for PM_{10}	-0.019	0.135	90.0

Firstly, both single and multi-pollutant models are essentially producing unbiased estimates. Secondly, the RMSPE values indicate that the multi-pollutant model outperforms the single pollutant model, with improvements of 14% and 16% respectively for NO_2 and PM_{10}. This is mainly because the correlation between NO_2 and PM_{10} of 0.627 (seen in Table 5.10) is substantial and hence improves the prediction. Thus knowing

the value of one pollutant at a site increases the predictive accuracy for the other pollutant. Table 5.9 also shows that the RMSPE for NO_2 is much higher than for PM_{10}, which is mainly because the variance/uncertainty in NO_2 observations is higher than that for PM_{10}. For example the marginal error variance for PM_{10} is about 0.024 from both single and multi-pollutant models, which compares to about 0.067 for NO_2 (Table 5.10). Finally, Table 5.9 also shows that both single and multi-pollutant models have the same performance in terms of the 95% CI Coverage, all of which are good because they are close to their nominal 95% levels. Note that the coverage for PM_{10} is lower at 90%, which is likely because the validation study is only based on 30 sites and hence maybe unstable.

TABLE 5.10: Results from the single pollutant model (4.3) and the multipollutant model (5.1) for 2010.

Model	**Variable**	**Posterior mean**	**Posterior 95% CI**
Multi-pollutant	$\text{corr}(NO_2, PM_{10})$	0.627	(0.490, 0.730)
	$\sigma^2_{NO_2}$	0.068	(0.051, 0.091)
	$\sigma^2_{PM_{10}}$	0.025	(0.018, 0.033)
Single pollutant	$\sigma^2_{NO_2}$	0.067	(0.054, 0.083)
	$\sigma^2_{PM_{10}}$	0.023	(0.018, 0.030)

5.6 Application

The multi-pollutant model proposed in this chapter has been applied to a real data set from Scotland from 2006-2010, to predict multi-pollutant concentrations, which are then used for the study of multi-pollutant health effects in the next chapter. The pollutants include NO_2 and PM_{10} which have been described in Section 5.2, and both the measurements and the modelled data are on the log scale. The predictors for both pollutants for each year contains the site types (e.g. kerbside, roadside, rural, background), modelled data, and temperature. Inference for the multi-pollutant model is implemented within a Bayesian framework via McMC simulation, using a mixture of Gibbs sampling steps and Metropolis-Hastings moves. The results presented in this section are based on 30,000 posterior samples after 20,000 burn-in iterations.

5.6.1 Model fitting

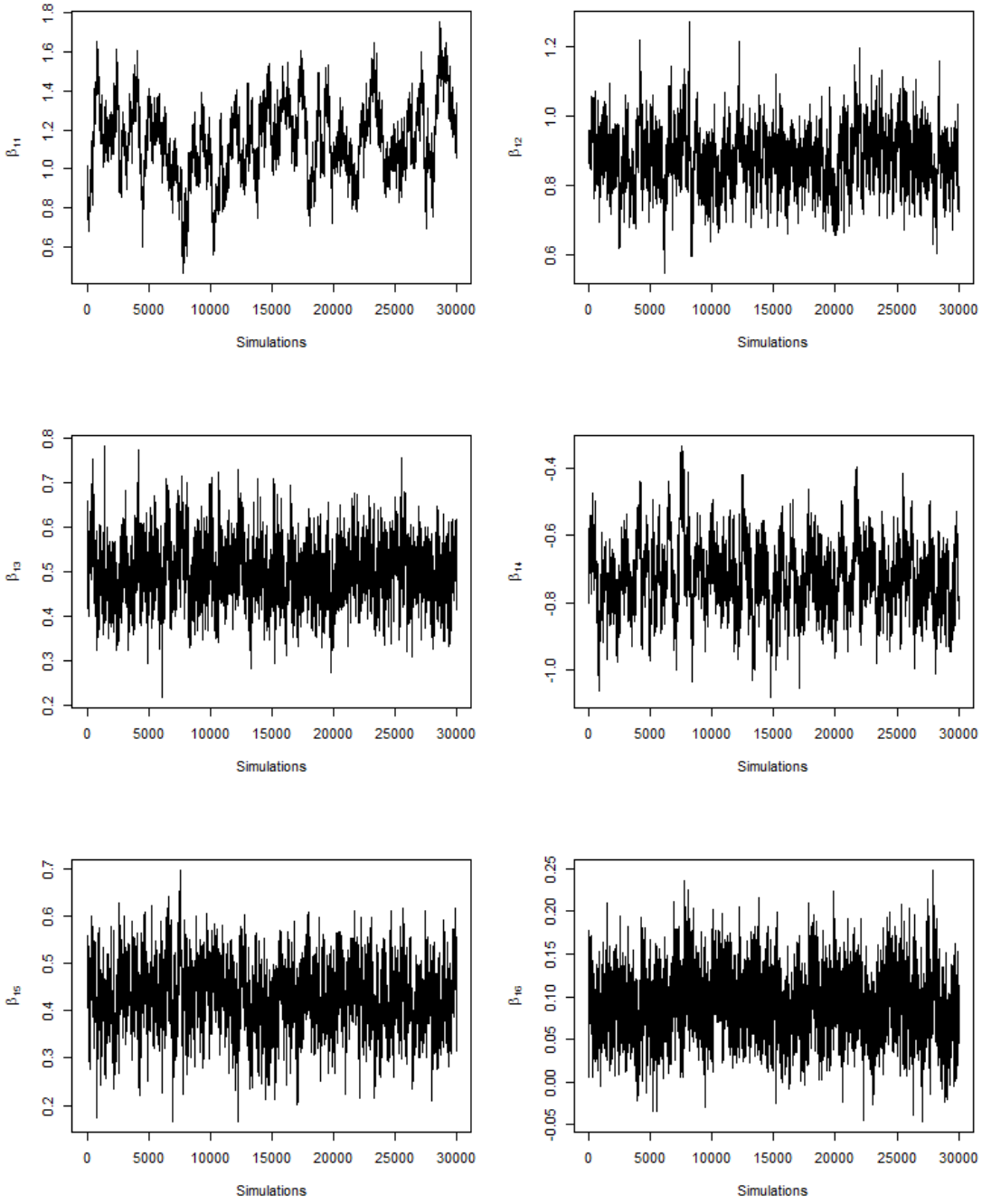
The multi-pollutant model was used to model the real data set in my study and the Markov chain Monte Carlo simulations for all the regression parameters had converged according to the Heidelberg and Welch Diagnostic with default arguments in R (eps=0.1, p-value=0.05) (Heidelberger and Welch [57], Heidelberger and Welch [58]). An example of the traceplot for the parameters of the simulations is shown in Figure 5.2 which displays the traceplot of the overall mean ($\boldsymbol{\beta}_1$) of the regression parameters.

The posterior means of the regression parameters from the application study are presented in Table 5.11, showing that all the regression parameters of both NO_2 and PM_{10} vary slightly over time, which is expected as the potential correlations between predictors and the response are likely unchanged within five years. This likely suggests that a simplification of model 5.2 by assuming constant regression parameters across time could be suitable for my research data. Table 5.11 also shows that the coefficient of modelled NO_2 ($\boldsymbol{\beta}_1$) data is much higher than for PM_{10} ($\boldsymbol{\beta}_2$), indicating the modelled NO_2 is a better predictor for measured NO_2 than the modelled PM_{10} for measured PM_{10}. This is not surprising as the correlation between modelled NO_2 and measured NO_2 is higher than that between modelled and measured PM_{10} (see the bottom right in Figure 5.1). Furthermore, it is also interesting to compare the output from the multi-pollutant model to those from the single pollutant model in the previous chapter. For example, the coefficient of the modelled data for NO_2 from the multi-pollutant model is lower than that from the single pollutant model (see Table 4.5). The drop of the coefficient is likely because in the multi-pollutant model the information on the correlation between NO_2 and PM_{10} helps to explain part of the variance in NO_2 response.

TABLE 5.11: Posterior means for the regression parameters from multi-pollutant model.

Parameter	$\boldsymbol{\beta}_1^{(1)}$	$\boldsymbol{\beta}_1^{(2)}$	$\boldsymbol{\beta}_1^{(3)}$	$\boldsymbol{\beta}_1^{(4)}$	$\boldsymbol{\beta}_1^{(5)}$	$\boldsymbol{\beta}_2^{(1)}$	$\boldsymbol{\beta}_2^{(2)}$	$\boldsymbol{\beta}_2^{(3)}$	$\boldsymbol{\beta}_2^{(4)}$	$\boldsymbol{\beta}_2^{(5)}$
Intercept	1.14	1.14	1.14	1.14	1.14	1.40	1.40	1.40	1.40	1.40
Kerbside	0.88	0.88	0.88	0.89	0.88	0.53	0.53	0.53	0.53	0.53
Roadside	0.51	0.50	0.49	0.49	0.49	0.29	0.30	0.30	0.30	0.31
Rural	-0.73	-0.73	-0.72	-0.72	-0.72	-0.31	-0.31	-0.31	-0.31	-0.31
Modelled data	0.43	0.43	0.43	0.44	0.44	0.29	0.28	0.29	0.29	0.30
Temperature	0.09	0.09	0.08	0.08	0.09	0.08	0.07	0.07	0.05	0.06

FIGURE 5.2: The results of the 30,000 McMC simulations for the overall mean of the regression parameters.



5.6.2 Model prediction

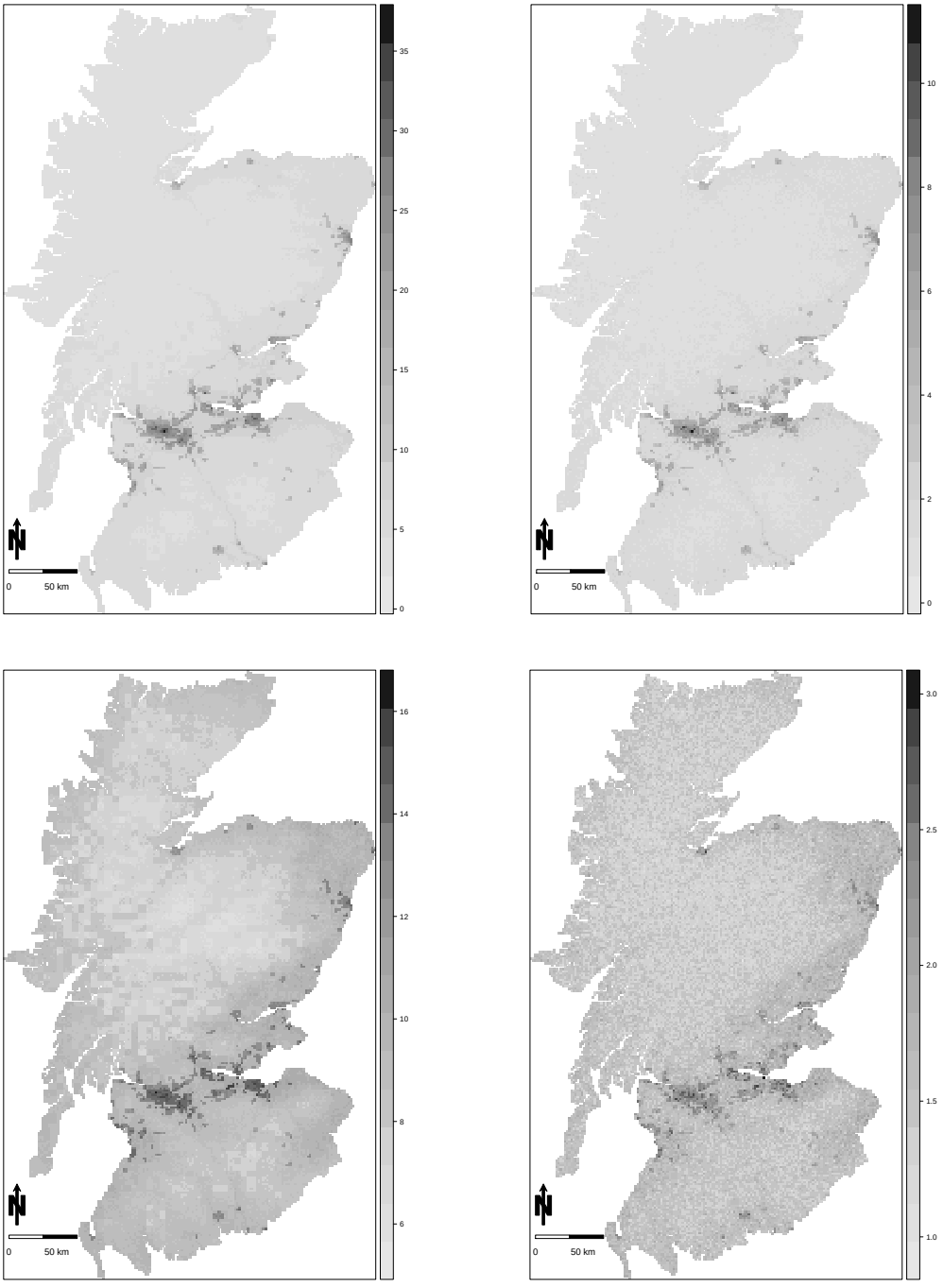
As the multi-pollutant model seems to outperform the single pollutant model (see Section 5.5), it is used to predict the pollution concentrations (NO_2 and PM_{10}) at 1 km resolution across mainland Scotland, which results in 68,448 prediction locations for each of $T = 5$ time periods (years). For a single location $\mathbf{s}_*$ and time period t, predictions of NO_2 $\left(X_1^{(t)}(\mathbf{s}_*)\right)$ and PM_{10} $\left(X_2^{(t)}(\mathbf{s}_*)\right)$ are made from the posterior predictive distribution $f\left(X_1^{(t)}(\mathbf{s}_*), X_2^{(t)}(\mathbf{s}_*)|\mathbf{X}\right)$, where $\mathbf{X}$ denotes the measured pollution data for both pollutants on the natural log scale for all time periods. $M = 100$ predictions are made from each posterior predictive distribution via composition sampling, sampling from the distribution $\mathrm{N}\left(\begin{bmatrix} \boldsymbol{Z}_1^{(t)}(\mathbf{s}_*)\boldsymbol{\beta}_1^{(t)} \\ \boldsymbol{Z}_2^{(t)}(\mathbf{s}_*)\boldsymbol{\beta}_2^{(t)} \end{bmatrix}, \sigma_t^2 \boldsymbol{C}_{2\times 2} \otimes \boldsymbol{I}_n\right)$, using the equation

$$\left[\left(X_1^{(t)}(\mathbf{s}_*)\right)^{(m)}, \left(X_2^{(t)}(\mathbf{s}_*)\right)^{(m)}\right] \mid \boldsymbol{\Theta}^{(m)} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{Z}_1^{(t)}(\mathbf{s}_*)(\boldsymbol{\beta}_1^{(t)})^{(m)} \\ \boldsymbol{Z}_2^{(t)}(\mathbf{s}_*)(\boldsymbol{\beta}_2^{(t)})^{(m)} \end{bmatrix}, (\sigma_t^2)^{(m)}(\boldsymbol{C}_{2\times 2})^{(m)} \otimes \boldsymbol{I}_n\right) \tag{5.5}$$

where $^{(m)}$ denotes the m^{th} McMC sample drawn from the posterior distribution of the model parameters and $\boldsymbol{Z}_1^{(t)}(\mathbf{s}_*)$, $\boldsymbol{Z}_2^{(t)}(\mathbf{s}_*)$ are the corresponding vectors of covariates for the prediction location $\mathbf{s}_*$ at time t for NO_2 and PM_{10}, respectively. The posterior mean and standard deviation of the M exponentiated predictions (as the measured data were modelled on the natural log scale) is taken at each grid point for NO_2, resulting in $Q = 68,448$ spatial point predictions $(\tilde{X}_1^{(t)}(\mathbf{s}_{1*}) \ldots, \tilde{X}_1^{(t)}(\mathbf{s}_{Q*}))$ and also their standard deviations for each of $T = 5$ time periods. The same procedure has been done for PM_{10}. For example, Figure 5.3 shows the maps of the predicted NO_2, PM_{10} and their predicted standard deviations on 1km grid across mainland Scotland for 2010, showing that the predicted NO_2, PM_{10} have a similar pattern to their modelled data (See Figure 4.1 and 5.1) where the concentrations are high in the lower middle part of the country. Figure 5.3 also shows that the predicted uncertainty tends to be high where the predicted concentrations are high. For 2006 to 2009, the maps of the predicted NO_2, PM_{10} are similar to 2010 and they are not shown here.

As the disease data used in the next chapter to investigate the multiple pollutants health effects relate to irregularly shaped geographical units (IGs), and are thus spatially misaligned to the grid level pollution predictions, I consider two different spatial aggregation

FIGURE 5.3: The predicted NO_2 and PM_{10} for 2010 from multi-pollutant model based on 1km resolution (unit: μgm^{-3}). Top left is the predicted NO_2 and top right is its predicted standard deviation, bottom left is the predicted PM_{10} and bottom right is its predicted standard deviation.



approaches here to convert the gridded data into the IG scale, the spatial mean and the spatial maximum value in each areal unit, which have been used in the previous chapter (equation 4.20). The areal level summaries are shown with an example of Figure 5.4, which displays the predicted NO_2, PM_{10} for 2010 on IG scale. The figure shows that the maps using spatial mean aggregation function are much smoother than those using spatial maximum aggregation function. This indicates the heterogeneous distribution of the gridded data within IGs.

5.7 Conclusion

In this chapter, a multi-pollutant model which extends the single pollutant model in the previous chapter is proposed, which allows the use of the correlation between pollutants to improve predictions. The performance of this multi-pollutant model is good since the simulation study shows that the model parameters are estimated without bias, the RMSE of each parameter is low, and the coverage of each parameter is quite close to its nominal 95% level. Furthermore, the validation study shows that the multi-pollutant model outperforms the single pollutant model proposed in the previous chapter in terms of the RMSE, with the improvements of 14% and 16% for NO_2 and PM_{10}, respectively. The multi-pollutant model is then applied to a real data set from Scotland from 2006-2010, and both the predicted NO_2 and PM_{10} concentrations for each 1km grid across Scotland are obtained. Finally, these gridded data are converted into IG scale by taking the spatial mean or maximum of the gridded data lying in each IG as the representative concentrations for that IG, which will be used in the next chapter to investigate the multiple pollutant health effects.

FIGURE 5.4: The predicted NO_2 and PM_{10} for 2010 from multi-pollutant model based on IG scale (unit: μgm^{-3}). Top left is based on using the max gridded NO_2 concentrations in each IG, top right is based on using the mean gridded NO_2 concentrations in each IG, bottom left is based on using the max gridded PM_{10} concentrations in each IG, bottom right is based on using the mean gridded PM_{10} concentrations in each IG.

Chapter 6

Health Effects of Exposure to Multiple Pollutants

6.1 Introduction

Recall that the pollutants people breathe contain a complex mixture of particle and gas phase pollutants (e.g. NO_2, PM_{10}, O_3, $PM_{2.5}$, SO_2, CO), therefore a traditional single pollutant health study in chapter 4 fails to account for the combined effects of pollutant mixtures. This chapter aims to explore the health effects of exposure to multiple pollutants simultaneously, and also to develop and compare several methods for incorporating exposure uncertainty into the investigation of health effects.

In the previous chapter a Bayesian multi-pollutant model allowing the correlation between pollutants to help improve the predictions was proposed, which extends the single pollutant model proposed in chapter 4. By applying this multi-pollutant model to the real data set in my study, 68448 1km gridded predictions for NO_2 and PM_{10} are obtained across mainland Scotland. Then the spatial mean and maximum aggregation functions are used to convert these grid level concentrations into areal units (IG), which results in two spatially representative concentration maps for each pollutant. Note that the pollutants considered in my study include only NO_2 and PM_{10} due to the sparse observations for the other pollutants.

A simple approach to investigate multi-pollutant exposure in epidemiologic analysis is to use a co-pollutant model, which fits a single regression model with multiple pollutants to

estimate the health effects of each pollutant (e.g. Yu et al. [144], Tolberta et al. [130]). However, a number of pollutants are highly correlated with each other such as $PM_{2.5}$ and NO_2 (Seaton and Dennekamp [116]), because they may be generated by common processes or be driven by similar factors such as meteorology, which means that it is inappropriate to include such pollutants in a single model as they are collinear, and thus their individual effects would not be well estimated.

Therefore, several statistical methods have been proposed to deal with this multi-collinearity issue. A simple approach is to construct an air quality indicator (AQI) based on the average of multiple pollutants, which is also called a *Score of Exposure*, with examples including Bruno and Cocchi [20], Powell and Lee [103], Lee et al. [77] and Hong et al. [61]. An AQI is a number that can be used by government agencies to inform the public how polluted the air currently is or how polluted its forecast to become. Different countries have their own air quality indices, corresponding to different national air quality standards. For example, a Daily Air Quality Index has been used in the UK, which depends on pollutant concentrations averaged over specified periods (*http://www.metoffice.gov.uk/*), while an Air Quality Health Index is adopted by Canada, which is calculated based on the relative risks of a combination of common air pollutants that are known to harm human health (*https://ec.gc.ca/*). Generally speaking, air pollution data are collected according to three dimensions: time, space and type of pollutant. These dimensions are reduced by means of aggregation processes, so as to obtain a synthetic value (Bruno and Cocchi [20]). However, the score is based on an *a-priori* attribution of weights to each pollutant (equal weights are usually taken), which ignores the correlation between the contributing pollutants and does not allow for determination of the effect of each type of pollutant. A special case of a *Score of Exposure* is to utilise a *Surrogate*, which is the use of the ambient concentration of one pollutant as an indicator of a combined exposure to several pollutants (Kim et al. [67]). Recently, another promising method to attribute the weights to each pollutant to obtain exposure metrics is based on health effects (risk-based metrics) (Oakes et al. [94]). The metrics incorporate health information and are primarily used to communicate the potential risk associated with air quality in a health study (To et al. [129]). There are two common ways to build risk-based metrics. The first is to weight the pollutant concentrations by scaling them to air quality standards developed to protect public health, and then sum

the scaled concentrations. The second is an additive combination of pollutant concentrations, in which the individual pollutant concentrations are weighted by their estimated effects on health. The risk-based metrics are based on determining the metric that has the strongest association with health (Pachon et al. [96]). The question regarding this metric is the appropriateness of using health information to create a metric that is then used as an explanatory variable in a health study.

Another approach, *Dimension reduction analysis*, can also be used to adjust for the issue of multicollinearity. Factor analysis (including Principal Components Analysis (PCA)) is a commonly used method of *Dimension reduction analysis*. It can minimize multicollinearity because the derived factors are orthogonal to one another, with examples including Rushworth et al. [109], Arif and Shah [3] and Qian et al. [105]. As factor analysis reduces the dimension of explanatory variables, it allows fewer potential parameterizations of main effects and interactions to be considered. However, there are some difficulties of factor analysis, which are the choice of the number of factors and the threshold loading value to interpret the factors.

A Bayesian hierarchical modelling approach can also be used to model multiple pollutants in which the pollutants are correlated, by adding a structured prior model for the exposure effects. MacLehose et al. [82] summarized four classes of higher-level (prior) distributions for incorporating similarities among multiple exposures, including two parametric and two semiparametric models. These models differ in how their prior distributions are specified. The first parametric model assumes the exposure effects of each pollutant follow a normal prior with fixed mean and variance which are determined by researchers. The lack of a prior distribution on the prior mean and variance is the distinguishing feature of this model. The second parametric model accounts for the uncertainty in the prior variance by placing a prior distribution on it. Rather than assuming a normal prior distribution, the two semiparametric models relax this assumption by letting the prior distribution be random. More details about these four models can be found in MacLehose et al. [82]. The drawbacks of this approach include the subjective information that many frequentists distrust, and the requirement of defining a complete statement of the *a-priori* distribution that many researchers find too exact to be realistic.

Recently, Bayesian kernel machine regression (BKMR) was also introduced by Bobb

et al. [17] as a new approach to study mixtures, in which the health outcome is regressed on a flexible function of the mixture (e.g. air pollution or toxic waste) components that is specified using a kernel function. Besides, Bayesian profile regression can also avoid the pitfalls of exposure variables that are highly collinear. It was used in recent studies (see e.g. Papathomas et al. [99], Pirani et al. [101] and Coker et al. [28]), which uses covariate values to observe joint patterns within the covariate data (reducing the dimensionality of the covariate data) and then relate health risks to joint patterns of exposure.

In epidemiological studies, exposure uncertainty is a key aspect because the exposures are only estimates and subject to uncertainty which needs to be accounted for. Blair et al. [16] concluded that exposure misclassification probably occurs in nearly every epidemiologic study. The effects of exposure uncertainty and the various methods proposed to correct for biases that result when exposure uncertainty is present have been discussed in Armstrong [4], Thomas et al. [127], Carroll et al. [24]. In addition, the effects of exposure uncertainty on point and interval estimates of exposure-disease associations have been investigated theoretically by Gladen and Rogan [48], Pickles [100], Brunekreef et al. [19], and Stram [123]. However, exposure uncertainty is still an under research topic in epidemiological study as there have been few original scientific publications that make use of existing methods for explicit exposure uncertainty correction in environmental or occupational health fields (Spiegelman [120]). Spiegelman [120] also suggested three reasons for this: the methods may be inappropriate for the particular features of environmental health studies, the use of methods to correct for exposure uncertainty requires exposure validation data which may not be available and the lack of human and technical capacity to perform the necessary adjustments. There have been a few recent studies considering exposure uncertainty while investigating air pollution health effects. For example, Dominici et al. [39] and Molitor et al. [90] developed Bayesian hierarchical measurement error models to incorporate exposure measurement error into the investigation of relative risk. Other examples include Bennett et al. [10], Shin et al. [118], Li et al. [81], Allodji et al. [2], and Kioumourtzoglou et al. [68].

In this chapter, both the single pollutant and the multi-pollutant health effects are investigated under the Bayesian framework. The single pollutant health effects are investigated based on the improved pollution predictions from the multi-pollutant model (5.1), while the multi-pollutant health effects are investigated by borrowing the theory of the added variable plot to handle the collinearity among multiple pollutants. I also

consider propagating the exposure error into the investigation of health effects. The reminder of this chapter is organised as follows. Section 6.2 provides the background to the study and a summary of the data. Section 6.3 outlines the modelling approaches used in this study. Section 6.4 presents the results of my study. Finally, Section 6.5 contains a concluding discussion.

6.2 Data description

The disease data and the covariates remain the same as chapter 4. The disease data are yearly numbers of admissions to non-psychiatric and non-obstetric hospitals in each IG from 2007 to 2011 with a primary diagnosis of respiratory disease, and the covariates include the percentage of people living in each IG who are in receipt of Job Seekers Allowance (JSA), and the median property price in each IG.

The pollutants considered in this study are annual mean NO_2 and PM_{10} from 2006-2010, which are the improved pollution predictions from the application of the multi-pollutant model proposed in chapter 5. These predictions consist of 100 sets to account for posterior uncertainty, with each set including predictions for each of NO_2 and PM_{10} on a 1km grid (68,448 in total) across mainland Scotland (see e.g. the maps of the predicted NO_2, PM_{10} and their predicted standard deviations on 1km grid across mainland Scotland for 2010 in Figure 5.3).

In this chapter, for all the models without considering exposure uncertainty, I use the mean of the 100 predictions for each grid to represent its predicted concentration. Then these grid level data are converted into areal units on which the disease data are aggregated, utilizing the spatial mean or maximum aggregation function. For each pollutant, this gives a predicted concentration for each spatial metric (spatial mean or maximum) for each IG.

For the models considering exposure error, each set of the grid level data are directly converted into areal units on which the disease data are aggregated, utilizing the spatial mean or maximum aggregation function. For each pollutant, this gives 100 predicted concentrations for each spatial metric (spatial mean or maximum) for each IG. The variance of these different predictions in each IG accounts for the uncertainty of exposure.

6.3 Methodology

6.3.1 Single pollutant disease model

Recall from chapter 4 that (Y_{kt}, E_{kt}) are the observed and expected numbers of disease cases in areal unit k during time period t, and the model presented here relates the pollution metrics in section (5.6.2) to the disease counts whilst accounting for other covariate factors and spatio-temporal autocorrelation. The model used to investigate single pollutant health effects is the same as chapter 4, and is given by:

$$\begin{aligned}
Y_{kt} \mid E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\
\ln(R_{kt}) &= \boldsymbol{b}_{kt}^T\boldsymbol{\alpha} + X_{qk(t-1)}\lambda + \phi_{kt}, \\
\boldsymbol{\alpha} &\sim \text{N}\left(\mathbf{0}, 1000\boldsymbol{I}\right), \\
\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1} &\sim \text{N}\left(\gamma\boldsymbol{\phi}_{t-1}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right), t = 2, ..., T, \\
\boldsymbol{\phi}_1 &\sim \text{N}\left(\mathbf{0}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right), \\
\lambda &\sim \text{N}\left(0, 1000\right), \\
\nu^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001), \\
\gamma, \rho &\sim \text{U}[0, 1],
\end{aligned} \tag{6.1}$$

where X_{qkt} is the spatially representative pollution concentration of pollutant q in areal unit k during year t. More details about the model can be seen in section 4.3.2.

6.3.2 Co-pollutant disease model

The health effects of exposure to multiple pollutants (2 pollutants in my study) can be investigated by simply applying Model (6.1) with multiple pollutants to estimate the health effects of each pollutant. This can be achieved by simply replacing the equation in Model (6.1) with:

$$\begin{aligned} \ln(R_{kt}) &= \boldsymbol{b}_{kt}^{T}\boldsymbol{\alpha} + X_{1k(t-1)}\lambda_1 + X_{2k(t-1)}\lambda_2 + \phi_{kt}, \\ \lambda_1, \lambda_2 &\sim \mathrm{N}\left(0, 1000\right), \end{aligned}$$

The co-pollutant disease model is a benchmark method to deal with multiple pollutants, however, this method ignores the high correlation between pollutants (X_{1kt} and X_{2kt}) and the health effects of each pollutant are likely to be poorly estimated. Therefore, I propose another multi-pollutant disease model which is given in the next section.

6.3.3 Multi-pollutant disease model

As the pollutants in my study (NO_2 and PM_{10}) are highly correlated (see chapter 5), the co-pollutant disease model does not provide reliable estimate of the health effects of each pollutant separately. Therefore, in this section I borrow the idea of the added variable plot to deal with this multicollinearity issue allowing both NO_2 and PM_{10} to be included in a disease model.

The added variable plot is also referred to as partial regression plot which is a commonly used method for obtaining a graphical evaluation of the effect of adding an explanatory variable (say, X_i) to model. An added variable plot illustrates the incremental effect on the response of specific terms by removing the effects of all other terms. It is formed by: (1) Compute the residuals of regressing the response variable against the explanatory variables but omitting X_i; (2) Compute the residuals from regressing X_i against the remaining explanatory variables; (3) Plot the residuals from (1) against the residuals from (2). If there is a pattern in this plot, then the adding explanatory variable is suggested to be added into the model (Ryan [110]).

The added variable plot is based on the fact that the residuals from a standard linear regression are uncorrelated with an explanatory variable in that model. Therefore, in my study, I regress one pollutant against another, e.g. regressing PM_{10} against NO_2. Then I take the residuals of this model, which are uncorrelated with NO_2 and represent the remaining signal of PM_{10} which cannot be explained by NO_2. Finally,

these residuals and the NO_2 data can be included in a single disease model without causing any multicollinearity issues, so as to investigate the health effects of exposure to both NO_2 and PM_{10} simultaneously.

Therefore, I firstly propose a temporally-varying linear model to regress one pollutant against another, which allows the intercept and slope for each time period $t = 1, ...T$ to be different. The reason why I use a temporally-varying linear model here is because both the intercept and slope vary for different time periods, and a simple linear model which forces both the intercept and slope to be the same for the entire time period will lead to an obvious pattern in the model residuals. The proposed temporally-varying linear model is given as,

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{X}_2^{(1)} \\ \boldsymbol{X}_2^{(2)} \\ ... \\ \boldsymbol{X}_2^{(T)} \end{bmatrix} &= \begin{bmatrix} \beta_0^{(1)}\mathbf{1} + \beta_1^{(1)}\boldsymbol{X}_1^{(1)} \\ \beta_0^{(2)}\mathbf{1} + \beta_1^{(2)}\boldsymbol{X}_1^{(2)} \\ ... \\ \beta_0^{(T)}\mathbf{1} + \beta_1^{(T)}\boldsymbol{X}_1^{(T)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \\ ... \\ \boldsymbol{\epsilon}^{(T)} \end{bmatrix} \\
\begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \\ ... \\ \boldsymbol{\epsilon}^{(T)} \end{bmatrix} &\sim \mathrm{N}(\mathbf{0}, \sigma^2 I_{nT \times nT})
\end{aligned} \tag{6.2}$$

where $\boldsymbol{X}_2^{(1)} = \{X_k : k = 1, ..., n\}$ is a vector of the pollution data across all IGs for pollutant 2 at time 1, $\boldsymbol{\epsilon}^{(1)} = \{\epsilon_k : k = 1, ..., n\}$ is a vector of the model residuals across all IGs for time 1, $\mathbf{1} = (1, ..., 1)_{n \times 1}$. Similarly, $\boldsymbol{X}_1^{(1)} = \{X_k : k = 1, ..., n\}$ is a vector of the pollution data across all IGs for pollutant 1 at time 1. This temporally-varying linear model is fitted using least-squares. Note that the purpose to use added variable plot idea here is to try to understand the additional effect of one pollutant given the other one, so I do not consider the other covariates (e.g. JSA).

As mentioned earlier, it is well known that the residuals from a linear regression are uncorrelated to a model explanatory variable. However, for the temporally-varying linear model proposed (model (6.2)), this statement is not so obvious, so I provide a proof as follows.

Theorem 6.1

Denote the estimated model residuals,

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}^{(1)} \\ \hat{\boldsymbol{\epsilon}}^{(2)} \\ ... \\ \hat{\boldsymbol{\epsilon}}^{(T)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_2^{(1)} - \hat{\beta}_0^{(1)}\mathbf{1} - \hat{\beta}_1^{(1)}\boldsymbol{X}_1^{(1)} \\ \boldsymbol{X}_2^{(2)} - \hat{\beta}_0^{(2)}\mathbf{1} - \hat{\beta}_1^{(2)}\boldsymbol{X}_1^{(2)} \\ ... \\ \boldsymbol{X}_2^{(T)} - \hat{\beta}_0^{(T)}\mathbf{1} - \hat{\beta}_1^{(T)}\boldsymbol{X}_1^{(T)} \end{bmatrix}$$

where $(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)})$ are the estimates minimizing the sum of squared errors $SSE(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)}) = \Sigma_{t=1}^{T}(\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)})^\top(\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)})$,

then

$\left(\hat{\boldsymbol{\epsilon}}^{(1)}, \hat{\boldsymbol{\epsilon}}^{(2)}, ..., \hat{\boldsymbol{\epsilon}}^{(T)}\right)^\top$ is uncorrelated to $\left(\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_1^{(T)}\right)^\top$

Proof 6.1

As $(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)})$ are estimated by minimizing the sum of squared errors $SSE(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)}) = \Sigma_{t=1}^{T}(\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)})^\top(\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)})$, set the partial derivatives of $SSE(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)})$ with respect to $\hat{\beta}_0^{(t)}$ and $\hat{\beta}_1^{(t)}$ equal to zero,

$$\begin{aligned} \frac{\partial SSE(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)})}{\partial \hat{\beta}_0^{(t)}} &= 0, \\ \Rightarrow -2 \times \mathbf{1} \cdot (\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)}) &= 0, \\ \Rightarrow \mathbf{1} \cdot \hat{\boldsymbol{\epsilon}}^{(t)} &= 0, \end{aligned} \tag{6.3}$$

$$\begin{aligned} \frac{\partial SSE(\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)})}{\partial \hat{\beta}_1^{(t)}} &= 0, \\ \Rightarrow -2 \times \boldsymbol{X}_1^{(t)} \cdot (\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)}) &= 0, \\ \Rightarrow \boldsymbol{X}_1^{(t)} \cdot \hat{\boldsymbol{\epsilon}}^{(t)} &= 0. \end{aligned} \tag{6.4}$$

Equation (6.3) and (6.4) give $(\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1}) \cdot \hat{\boldsymbol{\epsilon}}^{(t)} = 0$, where $\bar{X}_1 = \frac{\Sigma_{k=1}^{n}\Sigma_{t=1}^{T}X_{1kt}}{kt}$.

As

$$\mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)}) = \mathrm{E}(\boldsymbol{X}_2^{(t)} - \hat{\beta}_0^{(t)}\mathbf{1} - \hat{\beta}_1^{(t)}\boldsymbol{X}_1^{(t)}) = (\beta_0^{(t)}\mathbf{1} + \beta_1^{(t)}\boldsymbol{X}_1^{(t)} - \beta_0^{(t)}\mathbf{1} - \beta_1^{(t)}\boldsymbol{X}_1^{(t)}) = \mathbf{0},$$

we have

$$(\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1}) \cdot (\hat{\boldsymbol{\epsilon}}^{(t)} - \mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)})) = 0.$$

Therefore,

$$\Sigma_{t=1}^{T}(\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1}) \cdot (\hat{\boldsymbol{\epsilon}}^{(t)} - \mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)})) = 0.$$

That is,

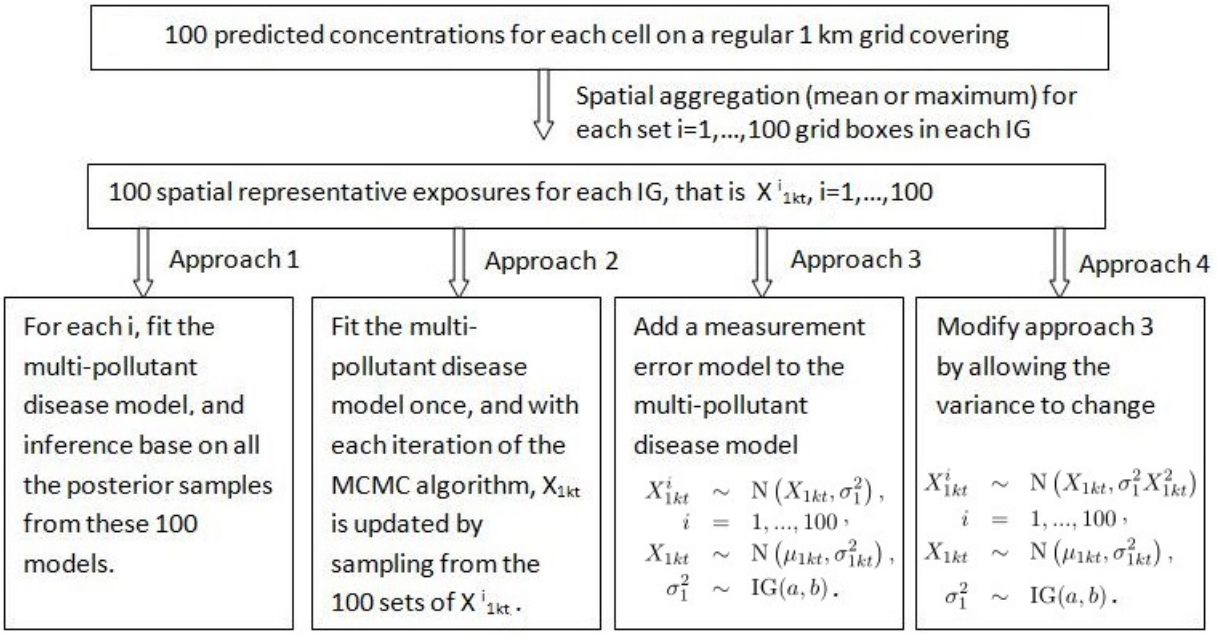
$$\begin{aligned} & \mathrm{cor}\left(\left(\hat{\boldsymbol{\epsilon}}^{(1)}, \hat{\boldsymbol{\epsilon}}^{(2)}, ..., \hat{\boldsymbol{\epsilon}}^{(T)}\right)^{\top}, \left(\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_1^{(T)}\right)^{\top}\right) \\ = & \frac{\Sigma_{t=1}^{T}(\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1}) \cdot (\hat{\boldsymbol{\epsilon}}^{(t)} - \mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)}))}{\sqrt{\Sigma_{t=1}^{T}(\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1}) \cdot (\boldsymbol{X}_1^{(t)} - \bar{X}_1\mathbf{1})\Sigma_{t=1}^{T}(\hat{\boldsymbol{\epsilon}}^{(t)} - \mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)})) \cdot (\hat{\boldsymbol{\epsilon}}^{(t)} - \mathrm{E}(\hat{\boldsymbol{\epsilon}}^{(t)}))}} \\ = & 0. \end{aligned}$$

End of proof

Until now, it has been shown that the residuals from the proposed temporally-varying linear model (6.2) are uncorrelated to the explanatory variable in the model which is pollutant $\boldsymbol{X}_1 = (\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_1^{(T)})$. Therefore, these model residuals $\hat{\boldsymbol{\epsilon}}$ and pollutant $\boldsymbol{X}_1$ can be put in a single disease model without causing multicollinearity issues. In other words, the information of both pollutants can be used simultaneously to investigate the multi-pollutant health effects. The multi-pollutant disease model can be obtained by simply replacing the equation in Model (6.1) by:

$$\begin{aligned} \ln(R_{kt}) &= \boldsymbol{b}_{kt}^{T}\boldsymbol{\alpha} + X_{1k(t-1)}\lambda + \hat{\epsilon}_{k(t-1)}\lambda_r + \phi_{kt}, \\ \lambda, \lambda_r &\sim \mathrm{N}(0, 1000), \end{aligned} \tag{6.5}$$

FIGURE 6.1: Four approaches to adjust for exposure uncertainty: X_{1kt} could be NO_2 or PM_{10}.



where $\hat{\epsilon_{kt}} = X_{2kt} - \hat{\beta}_0^{(t)} - \hat{\beta}_1^{(t)} X_{1kt}$ is the kth element of $\hat{\boldsymbol{\epsilon}}^{(t)}$ interpreted as the remaining signal from pollutant X_{2kt} which cannot be explained by pollutant X_{1kt}. λ_r is the regression coefficient for variable $\hat{\epsilon}_{kt}$ and assumed to be non-informative in the model by specifying a large variance in its prior distribution.

6.3.4 Dealing with exposure uncertainty

The multi-pollutant disease model (6.5) assumes the pollution data X_{1kt} is known and fixed in the disease model, which however is not true. Because X_{1kt} is based on the predicted concentrations from the multi-pollutant model, and the uncertainty of the predictions from the multi-pollutant model come from two main sources, with the first being the measurement error for the observed data and the second being the prediction uncertainty from the multi-pollutant model. Therefore, it is of interest to allow the uncertainty in the pollution predictions to be propagated through the model to investigate their impact on the health effects. In my study, I compare four approaches (see Figure 6.1) to allow the uncertainty of exposure to be propagated into the investigation of health effects for the multi-pollutant disease model.

Approach 1

The first approach to adjust for the exposure uncertainty is to fit the multi-pollutant disease model (6.5) 100 times, with exposure data being a different posterior sample from the multi-pollutant model each time. These different exposure data are the $M = 100$ sets of predictions from multi-pollutant model (see section 5.6.2), the variance of which represents the uncertainty of the exposure. Then model inference is obtained based on all the posterior samples from these 100 runs of the multi-pollutant disease model.

Approach 2

An alternative to propagate the exposure uncertainty is to incorporate it into the Bayesian hierarchical model by allowing the exposure to be sampled from the $M = 100$ sets of predictions (sample from the joint posterior) for each iteration of the McMC algorithm. That is the following model has been added to model (6.5)

$$\begin{aligned} X_{1kt} &= X^i_{1kt} \\ \hat{\epsilon}_{kt} &= \hat{\epsilon}^i_{kt} \\ i &= \text{Random(1,2,...,100)} \end{aligned} \tag{6.6}$$

where $\hat{\epsilon}^i_{kt}$ corresponds to the residuals resulting from fitting model (6.2) to the ith set of pollution predictions.

Approach 3

If the variation of the 100 exposure estimates within IGs are comparable to that across IGs, the performance of both approaches 1 and 2 could be poor. Specifically, the true effects between health risk and the pollution we want to estimate might be hidden by the massive posterior uncertainty in the pollution data. Given that the 100 exposure estimates within each IG are the estimates of a true exposure, their estimate errors belong to the classical type which assumes that the estimates are unbiased. Errors of the classical type arise when a quantity is measured by some device and repeated measurements vary around the true value (Heid et al. [56]). In this case, an alternative is to add a classical measurement error model as an extra level in the multi-pollutant disease model. This can be achieved by adding the following into model (6.5).

$$
\begin{aligned}
X^1_{1kt}, ..., X^{100}_{1kt} &\sim \mathrm{N}\left(X_{1kt}, \sigma_1^2\right), \\
X_{1kt} &\sim \mathrm{N}\left(\mu_{1kt}, \sigma^2_{1kt}\right), \\
\sigma_1^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001), \\
\hat{\epsilon}^1_{kt}, ..., \hat{\epsilon}^{100}_{kt} &\sim \mathrm{N}\left(\hat{\epsilon}_{kt}, \sigma_2^2\right), \\
\hat{\epsilon}_{kt} &\sim \mathrm{N}\left(\mu_{2kt}, \sigma^2_{2kt}\right), \\
\sigma_2^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001),
\end{aligned} \tag{6.7}
$$

where σ_1^2, σ_2^2 are the variances of X_{1kt}, $\hat{\epsilon}_{kt}$, respectively, which are allocated non-informative priors. In addition, weakly informative priors $\left(\mu_{1kt}, \sigma^2_{1kt}, \mu_{2kt}, \sigma^2_{2kt}\right)$ can be specified for X_{1kt} and $\hat{\epsilon}_{kt}$.

A few assumptions have been made for approach 3. In other words, this approach is appropriate when the predicted exposures (X^i_{1kt}) in each IG for each time period are unbiased, independent and from a normal distribution, and all the predicted IG exposures have a constant variance across IGs and time periods. These assumptions are also made for the residuals $(\hat{\epsilon}^i_{kt})$ from model (6.2), which is reasonable because the residuals are from a linear model (6.2) and the model assumptions are checked later (section 6.4.3). In the following, I justify these assumptions of X^i_{1kt}.

X^i_{1kt} are assumed to be unbiased, because the predicted concentrations from the multi-pollutant model proposed in chapter 5 appear to be unbiased (see Table 5.9). I check the independence of the predicted exposures in each IG using the Autocorrelation Function (ACF), the result of which shows that X^i_{1kt} are independent. For example, Figure 6.2 shows the ACF plots of X^i_{1kt} for a randomly selected (k, t). The dependence among X^i_{1kt} is because they come from the thinned McMC chain rather than the raw McMC chain. I check the normality assumption by a normal qq-plot. An example is shown in Figure 6.3, in which most of the points follow a linear trend suggesting the data can be treated as normally distributed.

The last assumption of approach 3 is the constant variance for X^i_{1kt} across IGs and time periods. Indeed, this is not a good assumption for my data, as the variances for

FIGURE 6.2: ACF plots of X^i_{1kt} for a randomly selected (k, t).

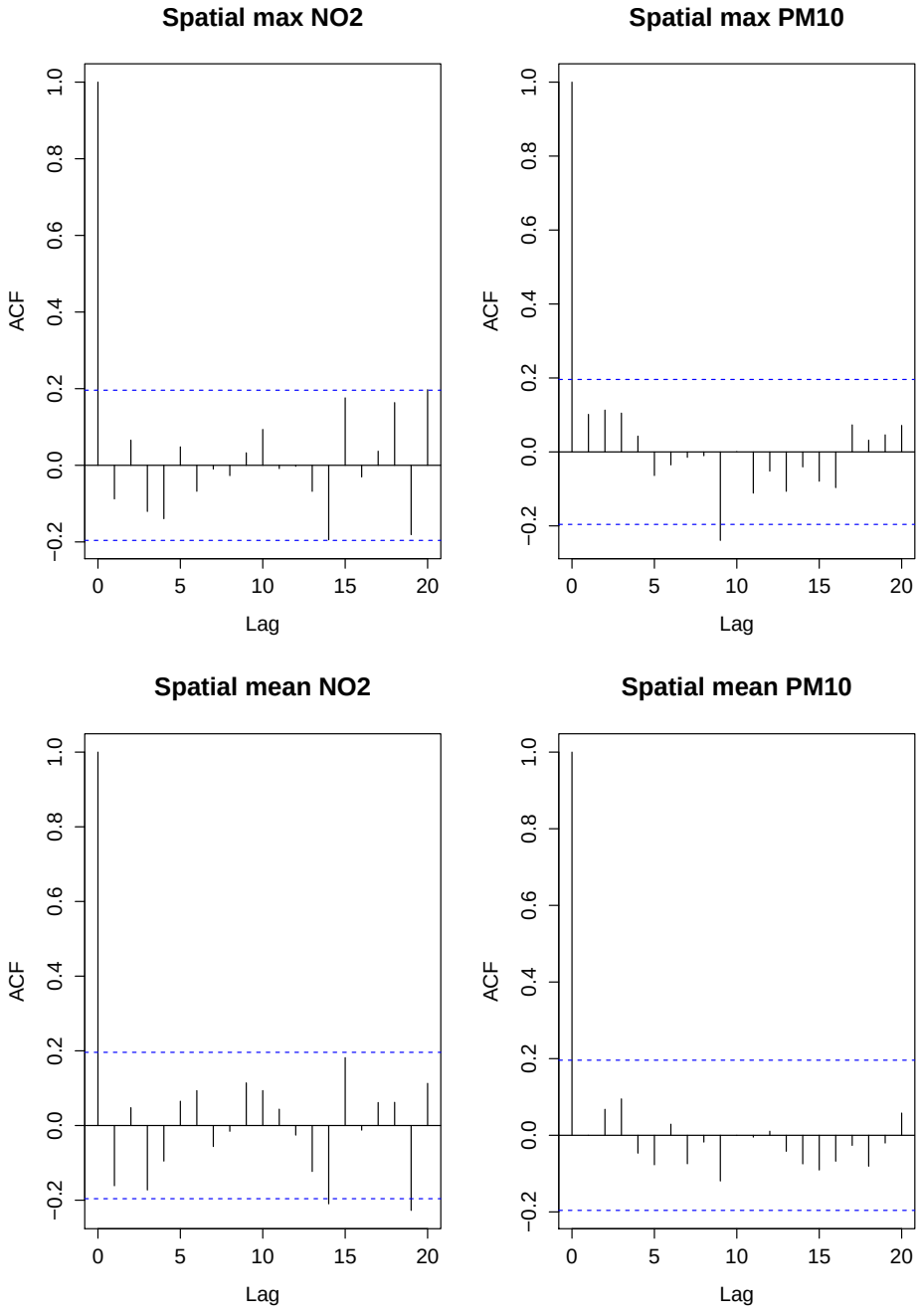


FIGURE 6.3: Normal qq plots of X^i_{1kt} for a randomly selected (k, t).

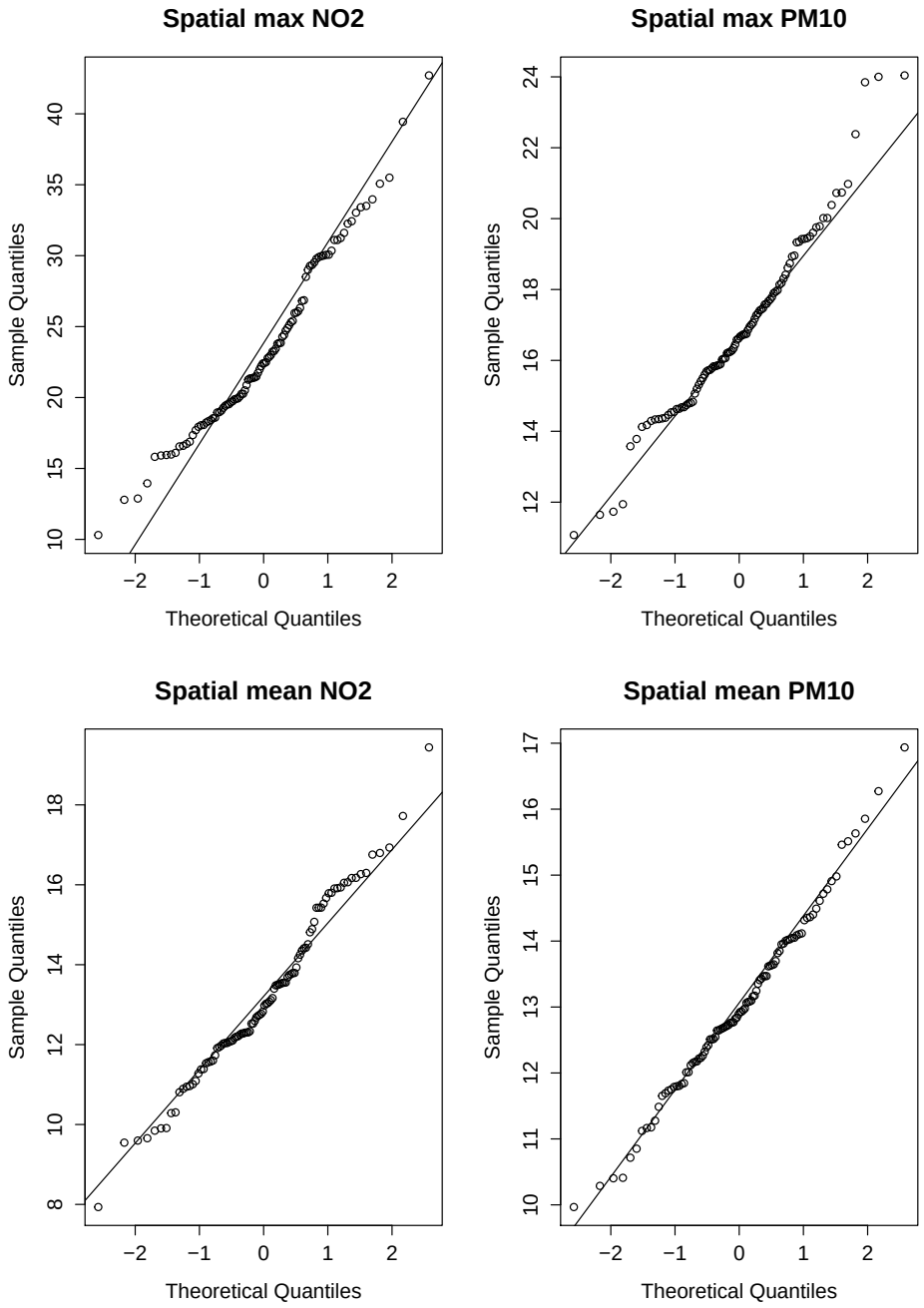
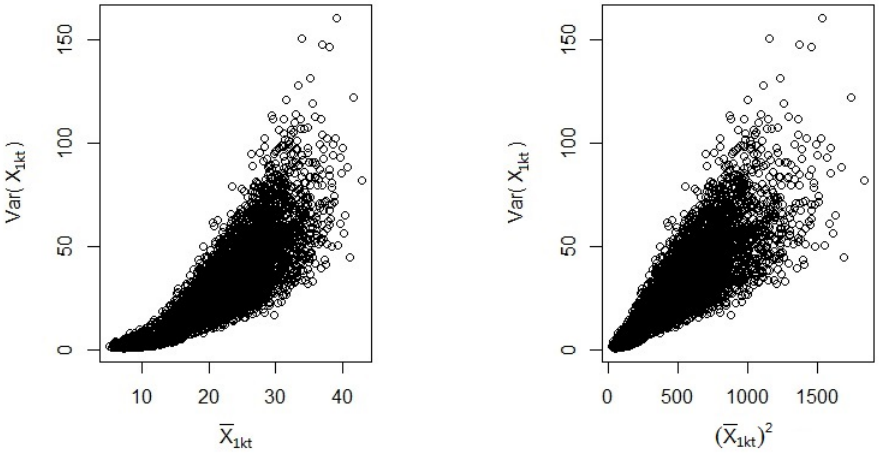


FIGURE 6.4: Scatter plots of the variance of X_{1kt} against $\bar{X}_{1kt}$ and $(\bar{X}_{1kt})^2$ for maximum NO_2.



X^i_{1kt} vary considerably (see Figure 6.6 where the distribution of the standard deviation of X_{1kt} within IGs is shown). However, this assumption makes the model much easier to fit and it is also interesting to investigate how this assumption affects the estimated health impact (compared to approach 4), which is the reason I consider it here.

Approach 4

This approach is to adjust the constant variance assumption of X_{1kt} in approach 3, which investigates the features of the variation of X_{1kt} first and then relax the assumption. I used the predictions of X_{1kt} to investigate the assumption and found that there is a linear trend between the posterior variance of X_{1kt} and $(\bar{X}_{1kt})^2$, where $\bar{X}_{1kt} = \frac{1}{100}\Sigma_{i=1}^{100} X^i_{1kt}$, $\text{var}(X_{1kt}) = \frac{1}{100-1}\Sigma_{i=1}^{100}(X^i_{1kt} - \bar{X}_{1kt})^2$. For example, Figure 6.4 shows the relationship between the variance of X_{1kt} and $(\bar{X}_{1kt})^2$ for spatial maximum NO_2.

Therefore, I adjust the constant variance assumption of X_{1kt} by allowing its variance to be different across the IGs and time periods. The model of X_{1kt} in approach 3 is extended to,

$$
\begin{aligned}
X^1_{1kt}, ..., X^{100}_{1kt} &\sim \mathrm{N}\left(X_{1kt}, \sigma_1^2 X^2_{1kt}\right), \\
X_{1kt} &\sim \mathrm{N}\left(\mu_{1kt}, \sigma^2_{1kt}\right), \\
\sigma_1^2 &\sim \text{Inverse-Gamma}(a = 0.001, b = 0.001),
\end{aligned} \tag{6.8}
$$

where the variation of X_{1kt} is assumed to be linearly dependent on X^2_{1kt} without an intercept term as suggested by Figure 6.4.

Note that it is under-researched how to allow the exposure variation in disease models, and in both approaches 3 and 4, this variation is allowed. The exposure has been specified a normal distribution prior which is what we believe about the pollution data before fitting the disease model, and the exposure is allowed to be updated depending on the health data.

6.4 Results

Inference for all models is implemented within a Bayesian framework via McMC simulation, using a mixture of Gibbs sampling steps and Metropolis-Hastings moves. The results from my study are based on 50,000 iterations (with 20,000 as burn-in iterations, after which the chain is checked for convergence). Model inference is obtained based on the remaining 30,000 posterior samples. The McMC simulation is implemented in R with the package CARBayesST (see Rushworth et al. [109]).

The regression parameters in disease models are presented as relative risks for a standard deviation increase in each covariate value, which are NO_2 6.84 μgm^{-3}, PM_{10} 1.872 μgm^{-3}, Logprice 0.38, JSA 2.35, residual standard deviation of mean PM_{10}, max PM_{10}, mean NO_2 and max NO_2 are 0.71 μgm^{-3}, 0.77 μgm^{-3}, 2.17 μgm^{-3}, 2.61 μgm^{-3}, respectively (see Table 6.3).

6.4.1 Single pollutant health effects

This section presents the long-term effects of each pollutant individually on health. The pollution data are the predicted concentrations from the multi-pollutant model proposed in the previous chapter, which contain two spatially representative pollution concentration metrics (spatial mean and maximum) for each of NO_2 and PM_{10}, and the health data are respiratory hospitalisation cases. The results are shown in Table 6.1.

TABLE 6.1: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the single pollutant disease model.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	0.997	1.030	1.017	1.056
	(0.983,1.010)	(1.013,1.048)	(1.005,1.029)	(1.046,1.067)
Logprice	0.920	0.920	0.923	0.915
	(0.910,0.931)	(0.910,0.931)	(0.913,0.934)	(0.906,0.926)
JSA	1.200	1.197	1.193	1.181
	(1.185,1.215)	(1.181,1.211)	(1.176,1.209)	(1.166,1.198)
ν^2	0.061	0.060	0.061	0.056
	(0.056,0.066)	(0.056,0.065)	(0.056,0.065)	(0.052,0.061)
ρ	0.925	0.900	0.870	0.683
	(0.886,0.955)	(0.856,0.940)	(0.785,0.926)	(0.593,0.773)
γ	0.832	0.829	0.829	0.814
	(0.796,0.869)	(0.792,0.865)	(0.792,0.864)	(0.778,0.851)
DIC	45120	45118	45117	45103

For NO_2, Table 6.1 shows that the spatial maximum NO_2 in each IG shows a significant relationship with respiratory disease while the spatial mean NO_2 does not. This is because the 95% credible interval of the relative risk for the former does not contain the neutral effect, 1, while the latter does. Specifically, Table 6.1 indicates that a 6.84 μgm^{-3} increase in peak NO_2 exposure is associated with 3% higher respiratory disease hospital admissions in Scotland, whereas no relationship is observed when the spatial mean NO_2 is used. This result is similar to what I found in chapter 4, where the pollution data are the DEFRA concentrations or the predicted pollution concentrations from the single pollutant model (4.3).

For PM_{10}, both the spatial mean and maximum PM_{10} in each IG show significant relationships with respiratory disease, indicating that a 1.872 μgm^{-3} increase in mean PM_{10} exposure is associated with 1.7% higher respiratory disease hospital admissions in Scotland, whereas it is 5.6% for a 1.872 μgm^{-3} increase in peak PM_{10} exposure. This

result is similar to what I had found in chapter 3, where the pollution data are the DEFRA data alone. Therefore, it validates the use of DEFRA PM_{10} data in my study.

Table 6.1 also shows that $\rho > 0.68$ and $\gamma > 0.8$ across the four models indicating high spatial and temporal autocorrelation in the disease data after the covariate effects have been accounted for. Note that ρ from model (Max PM_{10}) is much lower than the other models, indicating that the spatial autocorrelation in the disease data after the covariate effects have been accounted for in this model is lower. This is likely because the maximum metric for PM_{10} is able to capture more spatial correlation in the disease data, as the DIC for model (Max PM_{10}) is 45103 which is lower than those from the remaining models, 45120, 45118, 45117, respectively.

6.4.2 Co-pollutant health effects

The output of the co-pollutant disease model is displayed in Table 6.2, which indicates that the exposure in each IG using either the spatial mean or maximum PM_{10} has an adverse effect on health. However, those results for NO_2 show a beneficial effect, and this is unrealistic. This happens because NO_2 and PM_{10} are highly correlated. When both of them are fitted in a single model, the health effects for each pollutant are not well estimated. Table 6.2 also shows that the health effects of NO_2 are lower (even become beneficial) compared to those from the single pollutant health effects in Table 6.1, while those for PM_{10} are over estimated. In addition, the 95% credible interval for the relative risks are much wider than those from the single pollutant health effects (see Table 6.1), indicating that there is more uncertainty for the estimates of the relative risks. Therefore, the co-pollutant model is not suitable for investigating the health effects of correlated pollutants.

6.4.3 Multi-pollutant health effects

This section presents the results from fitting the multi-pollutant disease model. In my study, I consider two orders to combine both NO_2 and PM_{10} exposures into the disease model, with the first way is that $\boldsymbol{X}_1 = (\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_1^{(T)})$ represents NO_2, $\boldsymbol{X}_2 = (\boldsymbol{X}_2^{(1)}, ..., \boldsymbol{X}_2^{(T)})$ represents PM_{10}, while the other way is that $\boldsymbol{X}_1$ represents PM_{10}, $\boldsymbol{X}_2$ represents NO_2. In addition, there are two spatially representative concentrations for

TABLE 6.2: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the co-pollutant disease model.

Parameter	**Spatial mean**	**Spatial maximum**
NO_2	0.926	0.976
	(0.904,0.945)	(0.958,0.993)
PM_{10}	1.075	1.069
	(1.058,1.094)	(1.052,1.081)
Logprice	0.922	0.916
	(0.911,0.932)	(0.905,0.925)
JSA	1.190	1.182
	(1.174,1.206)	(1.168,1.199)
ν^2	0.058	0.056
	(0.054,0.063)	(0.052,0.061)
ρ	0.758	0.677
	(0.682,0.833)	(0.585,0.778)
γ	0.819	0.814
	(0.783,0.853)	(0.778,0.849)
DIC	45112	45100

the exposure. Therefore, in our study, there are four possible ways to fit model (6.2) and then model (6.5): either using spatial mean or maximum metrics, and where $\boldsymbol{X}_1$ represents either NO_2 or PM_{10}.

The model summaries for model (6.2) are shown in Table 6.3, in which the high R^2 values (from 0.845 to 0.941) indicate a high correlation between NO_2 and PM_{10}. The table also indicates that there is more variation among the model residuals from using the spatial maximum metric, compared to those using the spatial mean metric, because the R^2 from the former is lower than the latter for both NO_2 and PM_{10}.

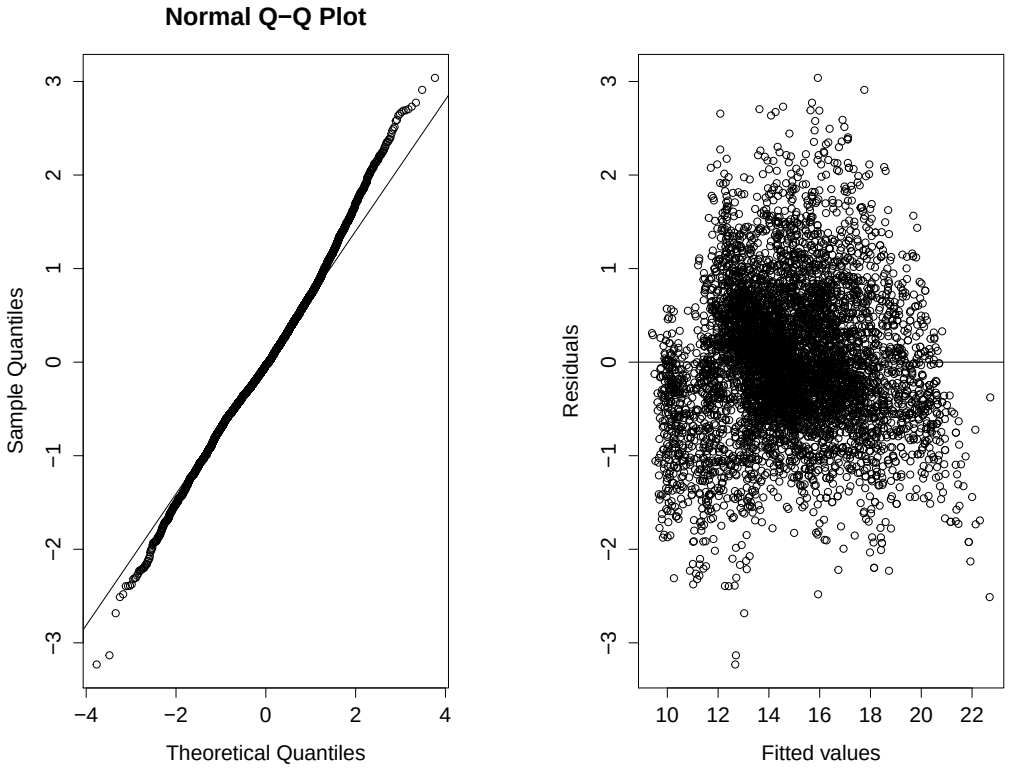
The fit of model (6.2) is assessed by checking the model residuals. There is no pattern in the residual plot against the fitted values, and the normal qq plot of the residuals shows that most of the points follow a linear trend suggesting the residuals of the model can be treated as normally distributed (given the size of data is 68448, the tails in the plot is acceptable). This indicates that the temporally-varying linear model is suitable for my data set. For example, Figure 6.5 displays the residual plot and its qq plot for regressing the spatial maximum PM_{10} against the spatial maximum NO_2.

After fitting model (6.2), the remaining signal of $\boldsymbol{X}_2$ which can not be explained by $\boldsymbol{X}_1$ is obtained, which is then added as an covariate into the multi-pollutant disease model. The results of the multi-pollutant disease model are presented in Table 6.4.

TABLE 6.3: Model summaries from model (6.2) (unit for residuals: μgm^{-3})

Residuals	$\mathbf{R^2}$	**sd(residuals)**
mean PM_{10}	0.941	0.71
max PM_{10}	0.908	0.77
mean NO_2	0.920	2.17
max NO_2	0.845	2.61

FIGURE 6.5: Model residuals from regressing spatial maximum PM_{10} against spatial maximum NO_2.



For models with NO_2, compared to the single pollutant health effects in Table 6.1, Table 6.4 shows that the remaining signal from PM_{10} does not affect the relative risk of NO_2. In addition, the DIC values do not change dramatically between the single pollutant disease model and the multi-pollutant disease model, indicating that the remaining signal from PM_{10} is not very helpful to explain the residuals from the disease data after accounting for NO_2 and other covariates. This is also supported by the non-significant relative risk of the remaining signal, whose 95% credible interval is (0.994,1.012) using the spatial mean metric and (0.997,1.013) using the spatial maximum metric. It likely suggests that NO_2 is sufficient to be used alone to investigate the health effects of exposure to both NO_2 and PM_{10} on respiratory disease in Scotland, given the high correlation between them.

TABLE 6.4: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	0.997	1.030	1.023	1.063
	(0.984,1.011)	(1.015,1.045)	(1.008,1.034)	(1.055,1.072)
Residuals PM_{10}	1.004	1.005	NA	NA
	(0.994,1.012)	(0.997,1.013)	NA	NA
Residuals NO_2	NA	NA	1.004	1.018
	NA	NA	(0.991,1.015)	(1.009,1.028)
Logprice	0.921	0.920	0.923	0.915
	(0.910,0.931)	(0.909,0.929)	(0.912,0.934)	(0.905,0.926)
JSA	1.199	1.196	1.189	1.176
	(1.183,1.217)	(1.178,1.214)	(1.174,1.204)	(1.160,1.193)
ν^2	0.061	0.060	0.061	0.055
	(0.056,0.065)	(0.056,0.065)	(0.056,0.065)	(0.051,0.059)
ρ	0.925	0.900	0.842	0.634
	(0.885,0.955)	(0.846,0.940)	(0.755,0.915)	(0.550,0.735)
γ	0.832	0.829	0.828	0.813
	(0.795,0.869)	(0.794,0.864)	(0.792,0.863)	(0.776,0.848)
DIC	45116	45126	45111	45098

For models with PM_{10}, Table 6.4 also shows that the remaining signal from spatial mean NO_2 which can not be explained by spatial mean PM_{10} is not significant in the disease model, given the PM_{10} pollutant and other covariates. However, by using the spatial maximum metric, the remaining signal from NO_2 is significant in the disease model, as the 95% credible interval for the relative risk does not contain the neutral effect, 1. This is likely because the residuals in model (Max PM_{10}) contain non-ignorable signals which can help to explain the disease data after accounting for PM_{10} and other covariates. This is consistent with Table 6.3, in which the R^2 for the temporally-varying linear model corresponding to model (Max PM_{10}) is the lowest, 0.845, indicating that there is more residual structure remaining.

As the components from PCA are also orthogonal, I compare the results from the multi-pollutant disease model to those by using PCA to deal with collinearity. For mean NO_2 and PM_{10}, the loadings for the first PC are both 0.707, while they are 0.707 and -0.707 for the second PC. For max NO_2 and PM_{10} the loadings for the first PC are both -0.707, while they are -0.707 and 0.707 for the second PC. The results based on PCA are shown in Table 6.5 which indicates that both components are significantly correlated to disease risk, as the 95% CI for each PC does not contain the neutral value, 1. For mean NO_2 and PM_{10}, the relative risk of PC1 is higher than 1 suggesting the adverse effects of the

combination of NO_2 and PM_{10} as the loadings for PC1 are positive. Similarly, for max NO_2 and PM_{10}, the relative risk of PC1 is lower than 1 which also suggests the adverse effects of the combination of NO_2 and PM_{10} as the loadings for PC1 are negative.

TABLE 6.5: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the PCA disease model (the relative risks are based on the same increasing units with those in Table 6.4).

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
PC1	1.135	0.730	1.035	0.918
	(1.041,1.249)	(0.680,0.800)	(1.011,1.063)	(0.900,0.941)
PC2	0.901	1.068	0.728	1.248
	(0.884,0.922)	(1.050,1.089)	(0.685,0.779)	(1.179,1.335)
Logprice	0.922	0.916	0.922	0.916
	(0.912,0.932)	(0.906,0.925)	(0.912,0.932)	(0.906,0.925)
JSA	1.189	1.185	1.189	1.185
	(1.172,1.204)	(1.170,1.199)	(1.172,1.204)	(1.170,1.199)
ν^2	0.058	0.056	0.058	0.056
	(0.054,0.063)	(0.051,0.060)	(0.054,0.063)	(0.051,0.060)
ρ	0.753	0.679	0.753	0.679
	(0.665,0.835)	(0.590,0.757)	(0.665,0.835)	(0.590,0.757)
γ	0.819	0.814	0.819	0.814
	(0.784,0.853)	(0.778,0.849)	(0.784,0.853)	(0.778,0.849)
DIC	45106	45099	45106	45099

In summary, the multi-pollutant disease model solves the multicollinearity issue from the highly correlated NO_2 and PM_{10}, which enable the investigation of the health effects of exposure to both pollutants simultaneously. The results from the multi-pollutant disease model (Table 6.4) indicate that a 6.84 μgm^{-3} increase in peak NO_2 exposure is associated with 3% higher respiratory disease hospital admissions in Scotland, whereas no relationship is observed when the spatial mean NO_2 is used, and a 1.872 μgm^{-3} increase in mean PM_{10} exposure is associated with 2.3% higher respiratory disease hospital admissions in Scotland, whereas it is 6.3% for a 1.872 μgm^{-3} increase in peak PM_{10} exposure. Table 6.4 also likely to suggest that there are independent health effects for NO_2 and PM_{10}, as the remaining signal from spatial max NO_2 which can not be explained by spatial max PM_{10} is still significantly associated with health.

6.4.4 Health effects with consideration of exposure uncertainty

This section investigates the multi-pollutant health effects while the exposure error is considered. Four approaches described in section 6.3.4 have been used to achieve this goal and the results are provided as follows.

6.4.4.1 Results from approach 1

The results from fitting the multi-pollutant disease model 100 times are given in Table 6.6, in which for each parameter, the point estimate and credible interval are calculated from 300,000 posterior samples (100 sets of the thinned posterior samples (3,000) from each fitting of the disease model).

TABLE 6.6: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 1.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	0.999	1.005	1.002	1.005
	(0.990,1.007)	(0.997,1.012)	(0.995,1.008)	(0.999,1.010)
Residuals PM_{10}	1.000	1.000	NA	NA
	(0.997,1.003)	(0.998,1.003)	NA	NA
Residuals NO_2	NA	NA	1.000	1.002
	NA	NA	(0.996,1.004)	(0.998,1.005)
Logprice	0.920	0.920	0.921	0.919
	(0.910,0.930)	(0.910,0.930)	(0.911,0.931)	(0.909,0.929)
JSA	1.199	1.198	1.198	1.198
	(1.183,1.215)	(1.182,1.215)	(1.182,1.214)	(1.181,1.214)
ν^2	0.061	0.061	0.061	0.061
	(0.056,0.065)	(0.056,0.065)	(0.056,0.065)	(0.056,0.065)
ρ	0.923	0.919	0.916	0.909
	(0.884,0.953)	(0.878,0.950)	(0.870,0.951)	(0.861,0.945)
γ	0.832	0.831	0.832	0.829
	(0.802,0.861)	(0.801,0.860)	(0.802,0.861)	(0.798,0.858)

Compared to the results from the multi-pollutant disease model without considering exposure uncertainty in Table 6.4, Table 6.6 shows that the relative risk using spatial maximum NO_2, spatial mean and maximum PM_{10} are no longer significant, and all the 95% CI of relative risk are narrower. According to my knowledge, the disappearance of the significant relative risk is likely not the real case. Instead, it is likely caused by the comparable variation of the 100 exposure estimates within IGs, compared to the exposure variation across IGs. Figure 6.6 shows the histograms for the standard deviation of the exposure within IGs for the spatial mean (maximum) of both NO_2 and PM_{10}, in which the red lines are the exposure variation across IGs while exposure uncertainty is not considered. These histograms show that the exposure estimate variations within IGs are comparable to those across IGs, as a non-ignorable part of the former is higher than the latter. The shrinkage of the relative risk CI in Table 6.6 is likely because the model is more certain about the neutral health effect of each exposure set (out of 100 sets).

FIGURE 6.6: Standard deviations (sd: μgm^{-3}) of the exposure within IGs: with the red lines are the sd of the exposure across IGs while exposure error is not considered.

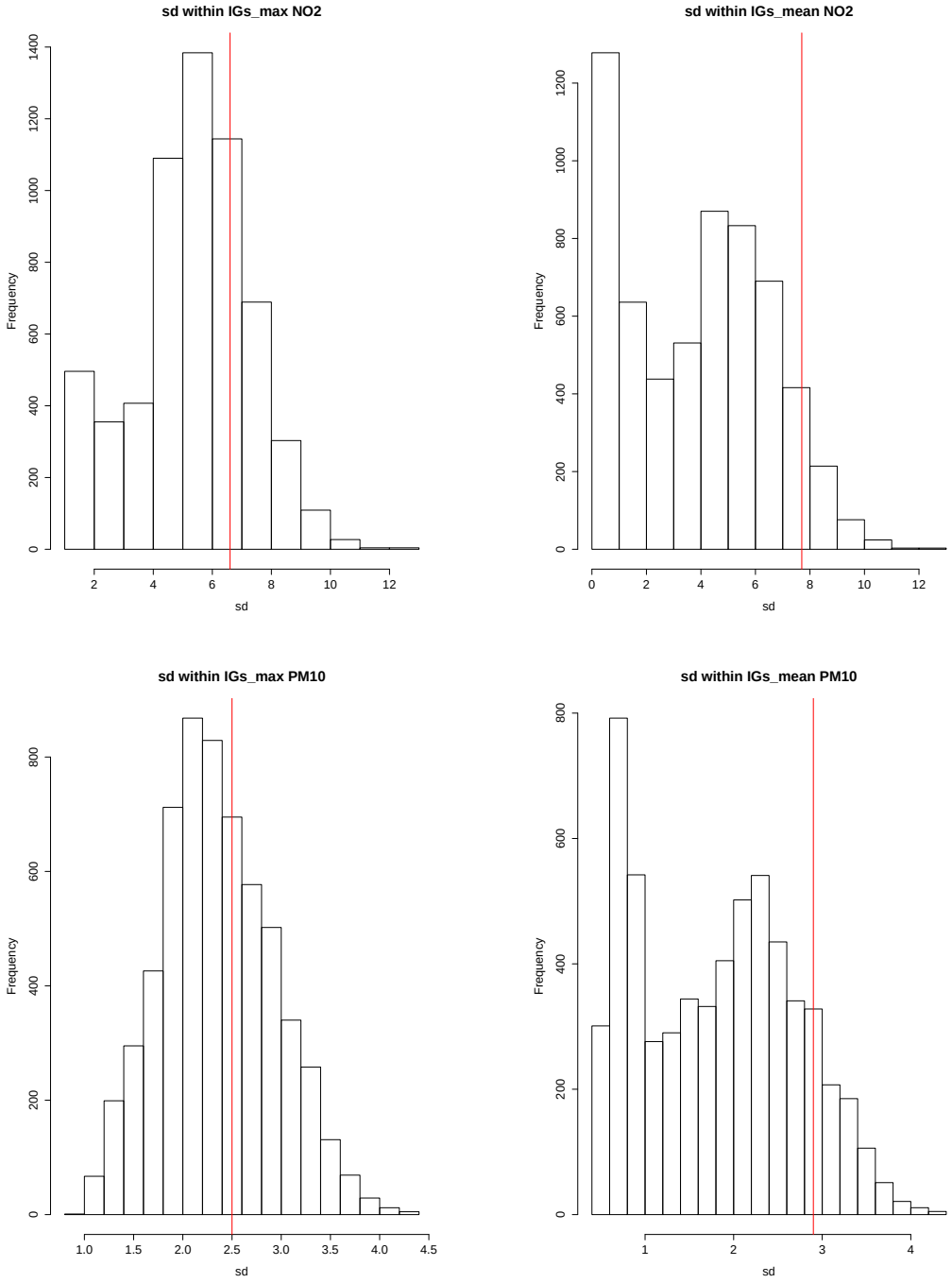


Table 6.6 also displays that the ρ for Model (Max NO_2, Mean PM_{10}, Max PM_{10}) are 0.919, 0.916,0.909, respectively. They are higher than those from the multi-pollutant disease model without considering exposure error (0.900, 0.842,0.634, see Table 6.4), indicating higher spatial auto-correlation of the model residuals while the exposure uncertainty is allowed. This is likely because the relative risks for these models in Table 6.6 are no longer significant and also lower than those from Table 6.4, indicating that the exposure in Table 6.4 helps explain more disease data variation and then there is less spatial structure left in the residuals.

6.4.4.2 Results from approach 2

While I consider the exposure uncertainty by allowing the exposure to be sampled from the $W = 100$ sets of predictions for each iteration of the McMC, the results are displayed in Table 6.7 which are similar to those from approach 1. That is the relative risk where using spatial maximum NO_2, spatial mean and maximum PM_{10} are no longer significant, and all the 95% CI become narrower. The disappearance of the significant relative risk is also likely because the exposure estimate variations within IGs are comparable with that across IGs. The spatial correlation parameters ρ are also higher compared to those models without considering exposure error (Table 6.4).

6.4.4.3 Results from approach 3

As the variation of the predicted exposures within IGs is comparable to that across the IGs and time periods, both approaches 1 and 2 are not appropriate for considering the exposure error in the investigation of the health effects of exposure. Approach 3, which considers the exposure error by adding a classical measurement error model into the multi-pollutant disease model, can address this issue because it allows the variation of exposure being estimated from the predicted exposures in each IG.

Before fitting approach 3, non-informative priors have been used for $X_{1kt}, \hat{\epsilon}_{kt}$, that is $\sigma^2_{1kt} = \sigma^2_{2kt} = 100,000$, $\mu_{2kt} = 0$ and μ_{1kt} is the average level of the exposure across IGs according to my knowledge (mean NO_2 17 μgm_{-3}, max NO_2 19 μgm_{-3}, mean PM_{10} 13 μgm_{-3}, max PM_{10} 15 μgm_{-3}). The convergence of $(X_{1kt}, \hat{\epsilon}_{kt}, \sigma^2_1, \sigma^2_2)$ in approach 3

TABLE 6.7: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the single pollutant disease model using approach 2.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	1.000	1.002	1.001	1.002
	(0.995,1.005)	(0.998,1.006)	(0.998,1.005)	(1.000,1.005)
Residuals PM_{10}	1.000	1.000	NA	NA
	(0.998,1.002)	(0.999,1.001)	NA	NA
Residuals NO_2	NA	NA	1.000	1.001
	NA	NA	(0.998,1.002)	(0.999,1.003)
Logprice	0.920	0.919	0.920	0.919
	(0.909,0.931)	(0.909,0.931)	(0.910,0.929)	(0.909,0.931)
JSA	1.198	1.198	1.198	1.196
	(1.183,1.215)	(1.182,1.215)	(1.182,1.213)	(1.181,1.214)
ν^2	0.061	0.061	0.061	0.061
	(0.057,0.065)	(0.057,0.065)	(0.056,0.065)	(0.056,0.065)
ρ	0.921	0.920	0.919	0.914
	(0.882,0.953)	(0.878,0.949)	(0.878,0.950)	(0.873,0.947)
γ	0.832	0.831	0.831	0.830
	(0.801,0.861)	(0.800,0.860)	(0.801,0.860)	(0.801,0.860)

was checked and observed. For example, Figure 6.7 shows the trace plot of σ_1^2, σ_2^2 and a randomly selected X_{1kt} and $\hat{\epsilon}_{kt}$ while applying approach 3 to spatial maximum PM_{10}, which indicates the parameters have converged.

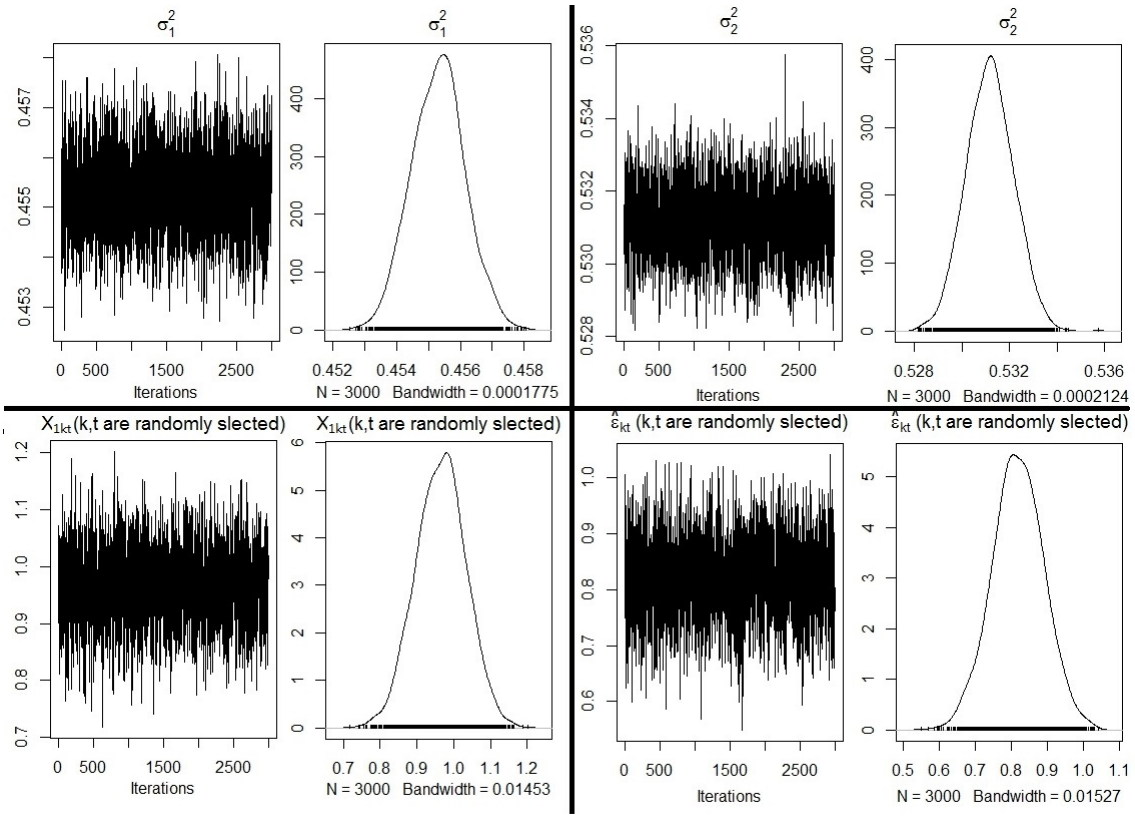
The posterior mean and 95% credible intervals for σ_1^2 (variance of X_{1kt} within an IG) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) are shown in Table 6.8. For all the models, σ_2^2 is higher than σ_1^2 and the CI is wider, indicating that there is less certain information about $\hat{\boldsymbol{\epsilon}}$ compared to $\boldsymbol{X}_1$.

TABLE 6.8: Posterior mean and 95% credible intervals for σ_1^2 (variance of X_{1kt} within an IG) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) from model (6.7).

X_{1kt}	$\sigma_1^2 = \mathbf{Var}(X_{1kt})$	$\hat{\epsilon}_{kt}$	$\sigma_2^2 = \mathbf{Var}(\hat{\epsilon}_{kt})$
Mean NO_2	0.260 (0.259, 0.261)	Residuals mean PM_{10}	0.825 (0.822, 0.828)
Max NO_2	0.415 (0.413, 0.416)	Residuals max PM_{10}	0.723 (0.721, 0.726)
Mean PM_{10}	0.310 (0.309, 0.311)	Residuals mean NO_2	0.715 (0.712, 0.717)
Max PM_{10}	0.455 (0.454, 0.457)	Residuals max NO_2	0.531 (0.529, 0.533)

The main results from approach 3 shown in Table 6.9 display that the relative risks of pollution in model (Max NO_2, Mean PM_{10}, Max PM_{10}) are significant, indicating that a 6.84 μgm^{-3} increase in peak NO_2 exposure is associated with 4% higher respiratory disease hospital admissions in Scotland, where as no relationship is observed when the

FIGURE 6.7: McMC trace plot for σ_1^2, σ_2^2 and a randomly selected $(X_{1kt},\ \hat{\epsilon}_{kt})$ from approach 3 by using spatial maximum of PM_{10}.



spatial mean NO_2 is used, and a 1.872 μgm^{-3} increase in mean PM_{10} exposure is associated with 1.7% higher respiratory disease hospital admissions in Scotland, whereas it is 4.7% for a 1.872 μgm^{-3} increase in peak PM_{10} exposure.

Similar to the results from the multi-pollutant disease model without considering exposure uncertainty (Table 6.4), Table 6.9 also shows that the remaining signal from PM_{10} is not helpful to explain the residuals from the disease data after accounting for NO_2 and other covariates. The 95% CIs of the relative risk contain the neutral effect, 1, which are (0.990,1.031) for using spatial mean metric and (0.984,1.010) for using spatial maximum metric. The same result has been found for the remaining signal from spatial mean NO_2, however, the remaining signal from spatial maximum NO_2 is significant in the model, the likely reason of this has been discussed in section 6.4.3.

Compared to Table 6.4, Table 6.9 also displays wider CI for the relative risk, which is to be expected, because the uncertainty of exposure $(\boldsymbol{X}_1)$ have been propagated into the estimation of the relative risk. The same is true for $\hat{\boldsymbol{\epsilon}}$.

TABLE 6.9: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 3.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	0.993	1.040	1.017	1.047
	(0.976,1.011)	(1.024,1.060)	(1.004,1.032)	(1.034,1.061)
Residuals PM_{10}	1.010	0.998	NA	NA
	(0.990,1.031)	(0.984,1.010)	NA	NA
Residuals NO_2	NA	NA	0.981	1.012
	NA	NA	(0.965,1.004)	(1.002,1.026)
Logprice	0.921	0.918	0.922	0.911
	(0.910,0.930)	(0.908,0.929)	(0.913,0.933)	(0.901,0.921)
JSA	1.200	1.197	1.194	1.188
	(1.184,1.218)	(1.182,1.214)	(1.176,1.210)	(1.170,1.204)
ν^2	0.061	0.060	0.061	0.059
	(0.056,0.065)	(0.056,0.065)	(0.057,0.065)	(0.055,0.063)
ρ	0.927	0.899	0.885	0.777
	(0.888,0.957)	(0.856,0.936)	(0.805,0.936)	(0.690,0.853)
γ	0.832	0.825	0.828	0.811
	(0.802,0.860)	(0.796,0.854)	(0.799,0.858)	(0.781,0.839)
DIC	45126	45128	45116	45116

6.4.4.4 Results from approach 4

The same as approach 3, before fitting approach 4, non-informative priors have been used for $X_{1kt}, \hat{\epsilon}_{kt}$. The convergence of $(X_{1kt}, \hat{\epsilon}_{kt}, \sigma_1^2, \sigma_2^2)$ in approach 4 was checked and observed. For example, Figure 6.8 shows the trace plot of σ_1^2, σ_2^2 and a randomly selected X_{1kt} and $\hat{\epsilon}_{kt}$ while applying approach 4 to spatial maximum PM_{10}, which indicates the model parameters have converged.

The posterior mean and 95% credible intervals for σ_1^2 (slope between $\text{var}(X_{1kt})$ and X_{1kt}^2) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) are shown in Table 6.10. σ_2^2 is expected to be the same with those from approach 3, as approach 4 only adjusts the constant variance for X_{1kt}. The values for σ_1^2 which is the slope between $\text{var}(X_{1kt})$ and X_{1kt}^2, indicate that the dependency between $\text{var}(X_{1kt})$ and X_{1kt}^2) for NO_2 is much stronger than PM_{10} as the slopes for NO_2 are steeper than those for PM_{10}.

The main results from approach 4 are shown in Table 6.11. According to this table, the relative risks of the pollutants (Max NO_2, Mean PM_{10}, Max PM_{10}) are significant, indicating that a 6.84 μgm^{-3} increase in peak NO_2 exposure is associated with 3.4% higher respiratory disease hospital admissions in Scotland, whereas no relationship is

FIGURE 6.8: McMC trace plot for σ_1^2, σ_2^2 and a randomly selected $(X_{1kt},\ \hat{\epsilon}_{kt})$ from approach 4 by using spatial maximum of PM_{10}.

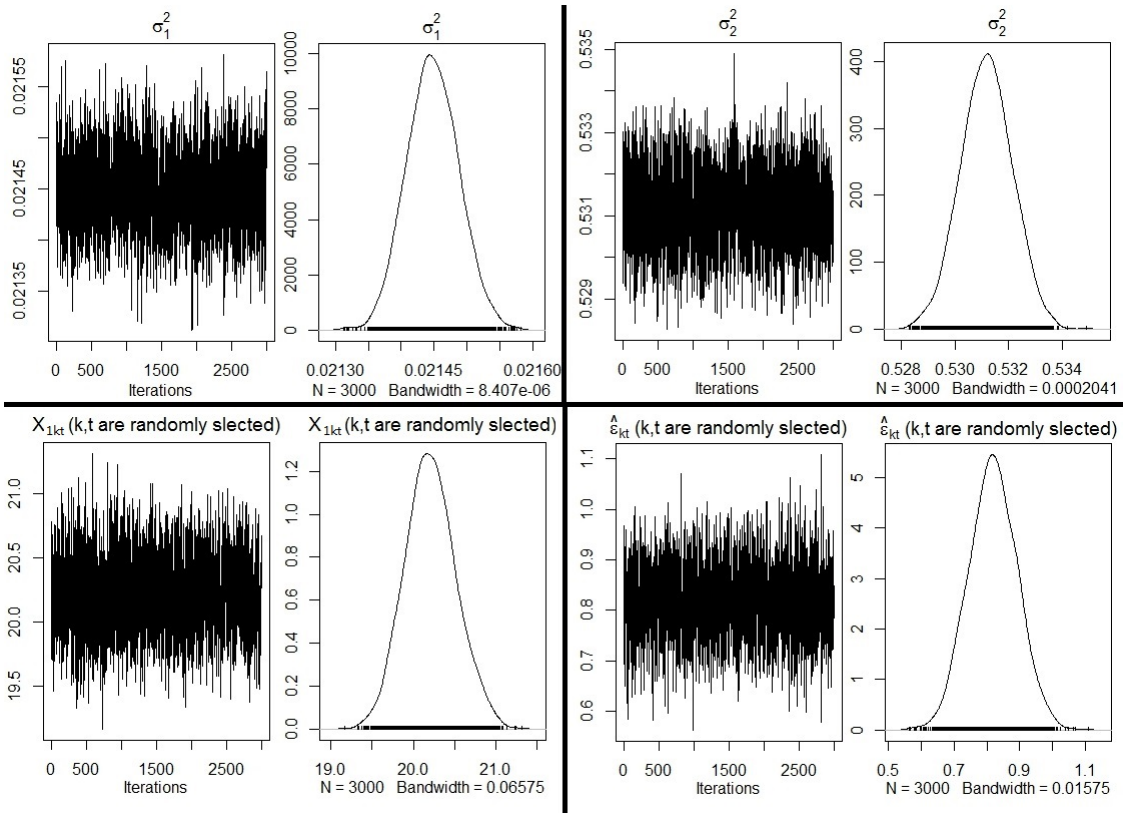


TABLE 6.10: Posterior mean and 95% credible intervals for σ_1^2 (slope between var(X_{1kt}) and X_{1kt}^2) and σ_2^2 (variance of $\hat{\epsilon}_{kt}$ within an IG) from model (6.8).

X_{1kt}	σ_1^2**(slope between var(X_{1kt}) and X_{1kt}^2)**	$\hat{\epsilon}_{kt}$	$\sigma_2^2 =$ **var**$(\hat{\epsilon}_{kt})$
Mean NO_2	0.0435 (0.0434, 0.0437)	Residuals mean PM_{10}	0.825 (0.822, 0.828)
Max NO_2	0.0568 (0.0566, 0.0570)	Residuals max PM_{10}	0.723 (0.721, 0.726)
Mean PM_{10}	0.0172 (0.0171, 0.0173)	Residuals mean NO_2	0.715 (0.712, 0.717)
Max PM_{10}	0.0214 (0.0214, 0.0215)	Residuals max NO_2	0.531 (0.529, 0.533)

observed when the spatial mean NO_2 is used, and a 1.872 μgm^{-3} increase in mean PM_{10} exposure is associated with 1.4% higher respiratory disease hospital admissions in Scotland, whereas it is 3.3% for a 1.872 μgm^{-3} increase in peak PM_{10} exposure.

As the only difference between approach 4 and 3 is that the former allows the variance of X_{1kt} to be different across IGs and time periods while the latter assumes a constant variance for X_{1kt} across k and t, I compare the relative risk of X_{1kt} from approach 4 to that from approach 3. Firstly, the point estimates of the relative risk of X_{1kt} from approach 4 are slightly lower than approach 3 but the statistical significance is consistent for both approaches. That is the relative risks of pollutant in model (Max NO_2, Mean PM_{10}, Max PM_{10}) are significant. Secondly, the credible intervals of the relative risk

of X_{1kt} from approach 4 are narrower than approach 3, indicating that allowing the varying variance of X_{1kt} can reduce the uncertainty of the estimates of the relative risk of X_{1kt}.

It is also of interest to compare the results from approach 4 to the multi-pollutant disease model which doesn't consider the uncertainty of exposure (Table 6.4). For model Mean NO_2, Mean PM_{10} and Max PM_{10}, the point estimates of the relative risk of X_{1kt} from approach 4 drop. For model Mean NO_2, Max NO_2 and Mean PM_{10}, the credible intervals of the relative risk of X_{1kt} from approach 4 are narrower even though this approach propagates the uncertainty of pollution X_{1kt} into the estimation of the relative risk. This is likely because the pollutant X_{1kt} in approach 4 is allowed to be updated depending on the health data which will probably be informative for X_{1kt}, then the uncertainty of the relative risk of X_{1kt} is reduced. Note that approach 3 also allows X_{1kt} to be updated depending on the health data, but the credible intervals of relative risk of X_{1kt} are still wider than those from the multi-pollutant disease model which doesn't consider the uncertainty of exposure (Table 6.4). This is likely because the assumption of the constant variance of X_{1kt} in approach 3 is not appropriate which weaken the improvement from the informative health data.

TABLE 6.11: Posterior means and 95% credible intervals of the regression, autocorrelation and variance parameters from fitting the multi-pollutant disease model using approach 4.

Parameter	**Mean NO_2**	**Max NO_2**	**Mean PM_{10}**	**Max PM_{10}**
Pollutant	0.992	1.034	1.014	1.033
	(0.979,1.002)	(1.021,1.046)	(1.003,1.024)	(1.024,1.043)
Residuals PM_{10}	1.013	0.998	NA	NA
	(0.992,1.032)	(0.985,1.009)	NA	NA
Residuals NO_2	NA	NA	0.978	1.012
	NA	NA	(0.954,1.004)	(1.004,1.024)
Logprice	0.920	0.918	0.922	0.912
	(0.909,0.931)	(0.908,0.929)	(0.912,0.931)	(0.901,0.921)
JSA	1.202	1.193	1.194	1.186
	(1.183,1.217)	(1.175,1.208)	(1.179,1.209)	(1.171,1.203)
ν^2	0.061	0.060	0.061	0.059
	(0.056,0.065)	(0.056,0.065)	(0.056,0.065)	(0.055,0.063)
ρ	0.930	0.889	0.885	0.778
	(0.894,0.959)	(0.835,0.931)	(0.822,0.932)	(0.689,0.850)
γ	0.832	0.825	0.829	0.811
	(0.802,0.862)	(0.795,0.854)	(0.799,0.858)	(0.781,0.841)
DIC	45128	45125	45120	45111

6.5 Discussion and conclusion

In this chapter I investigated the health effects from exposure to NO_2 and PM_{10} simultaneously, and also tried four approaches to propagate the exposure uncertainty into the investigation of health effects.

The investigation of the impact from the exposure to multiple pollutants is a natural extension to the single pollutant health effects (see chapter 4), because the polluted air people breathe is a mixture of different pollutants. In this study, I showed that the benchmark method, co-pollutant disease model, could lead to poor estimation of the influence of each individual pollutant, given NO_2 and PM_{10} are highly correlated (multicollinearity issue). Therefore, I proposed a temporally-varying linear model to regress one pollutant ($\boldsymbol{X}_2$) against another ($\boldsymbol{X}_1$), the residuals of which are then included along with ($\boldsymbol{X}_1$) into a single disease model to investigate the impact of exposure to both pollutants simultaneously thus resolving the multicollinearity issue.

The residuals from the temporally-varying linear model are interpreted as the remaining signal from ($\boldsymbol{X}_2$) which can not be explained by ($\boldsymbol{X}_1$), and the residual coefficient in the multi-pollutant disease model can be interpreted as the impact of the variation in one pollutant which can not be explained by another pollutant on the risk. Therefore, whether $\boldsymbol{X}_1$ (main pollutant in the multi-pollutant disease model) represents NO_2 or PM_{10} should depend on which pollutant we want to investigate. In this study, I consider both situations ($\boldsymbol{X}_1$ represents NO_2 or PM_{10}) and the results show that a 6.84 μgm^{-3} increase in peak NO_2 exposure is associated with 3.4% higher respiratory disease hospital admissions in Scotland, whereas no relationship is observed when the spatial mean NO_2 is used, and a 1.872 μgm^{-3} increase in mean PM_{10} exposure is associated with 1.4% higher respiratory disease hospital admissions in Scotland, whereas it is 3.3% for a 1.872 μgm^{-3} increase in peak PM_{10} exposure. The results also indicate that the remaining signal from both spatial mean and maximum PM_{10} which can not be explained by NO_2, are not helpful to explain the residuals from the disease data after accounting for NO_2 and other covariates. Similarly, the remaining signal from spatial mean NO_2 which can not be explained by spatial mean PM_{10} is not significant in the disease model, given the PM_{10} pollutant and other covariates. However, when using spatial maximum metric, the remaining signal from NO_2 is significant in the disease model. This is likely because the

residual variable in this model contains non-ignorable signal which can help to explain the disease data after accounting for the PM_{10} and other covariates.

Allowing the exposure uncertainty to be propagated into the investigation of health impact is important in epidemiological studies, because the the predicted exposures are likely to contain errors and uncertainties. In this study, I consider four approaches to adjust the exposure uncertainty. The first one is to fit the multi-pollutant disease model 100 times, with a different posterior predictive sample for pollution each time. The second approach is to incorporate the uncertainty of the exposure into the muti-pollutant disease model by allowing the exposure being sampled from the $M = 100$ sets of predictions for each iteration of the McMC algorithm. The third approach is to add a classical measurement error model as an extra level in multi-pollutant disease model, while the last approach is similar to the third approach except that the variance of $(\boldsymbol{X}_1)$ is allowed to be varied across IGs and time periods which matches the real data set better. Note that it is a two stage process to propagate exposure uncertainty in my study as the exposure predictions were obtained before fitting the disease models. However, a holistic approach by combining the exposure model and the disease model into a unified model could also be possible. This unified model will allow the exposure uncertainty to propagate directly into the disease models as the relative risks are estimated based on the pollution observations rather than the predictions from exposure model.

The results suggest that both approaches 1 and 2 appear to perform poorly as the significant relative risks observed previously have disappeared, due to the variation of the exposure estimates within IGs being comparable to that across IGs. In contrast, both approaches 3 and 4 maintain the significant relative risks while incorporating the exposure uncertainty. The results also show that the point estimates of the relative risk of X_{1kt} from approach 4 are slightly lower than approach 3 but the statistical significance is consistent for both approaches. That is the relative risks of pollutant in model (Max NO_2, Mean PM_{10}, Max PM_{10}) are significant. In addition, the credible intervals of the relative risk of X_{1kt} from approach 4 are narrower than approach 3, indicating that allowing the varying variance of X_{1kt} can reduce the uncertainty of the estimates of the relative risk of X_{1kt}. Compared to the multi-pollutant disease model which does not consider the uncertainty of exposure, the 95% credible intervals for relative risk of X_{1kt} from approach 4 are narrower even though this approach propagates the uncertainty of pollution X_{1kt} into the estimation of the relative risk. This is likely because the pollutant

X_{1kt} in approach 4 is allowed to be updated depending on the health data which will probably be informative for X_{1kt}, thus the uncertainty of the relative risk of X_{1kt} is reduced.

Note that both approaches 3 and 4 are computationally expensive because a large number of parameters (as the exposure for each IG and each period is treated as an unknown parameter) are updated in each iteration of the McMC algorithm. In my study, it takes only about 15 minutes (using R software) to fit the multi-pollutant disease model (6.5) without considering exposure uncertainty based on a normal PC. However, the computational burden dramatically increases while implementing either approach 3 or 4, which requires about 7 hours. Therefore, I wrote C++ subroutines to make the coding computationally efficient and the running time drops sharply from 7 hours to 35 minutes.

In this study, two spatial aggregation functions (mean and maximum) have been used to construct spatially representative pollution concentrations, while the majority of epidemiological studies use the average (mean) concentration (see e.g. Maheswaran et al. [84]; Lee et al. [76] and Warren et al. [136]). The results suggest that the choice of spatial aggregation metric used to quantify areal level pollution concentrations has a major impact on the resulting health effect estimate, which naturally leads to the question of which metric should one use. For example, spatial maximum metric for NO_2 is likely to be better than mean if the population are dense near the main roads, because NO_2 concentrations are usually higher near main roads where most of the exhaust fumes are produced.

Chapter 7

Conclusion

In this thesis, the long-term air pollution health effects had been investigated using a spatial ecological design in which the exposure uncertainty was incorporated into the estimation of health effects. The study region was mainland Scotland, UK, which consists of 1,207 Intermediate Geographies, each having an average population of around 4,300 people.

The disease data were yearly numbers of admissions to non-psychiatric and non-obstetric hospitals aggregated in each Intermediate Geography from 2007 to 2011 with a primary diagnosis of respiratory disease, while the pollutants contain NO_2 and PM_{10} reported as the annual mean and come from a sparse monitoring network and an atmospheric dispersion model (DEFRA). I used the pollution data from 2006 to 2010 rather than from 2007 to 2011, to make sure that the exposure occurred before the hospital admissions. In addition, other confounding variable were also considered in this study to describe the spatial pattern in disease risk, including the percentage of people living in each IG who are in receipt of Job Seekers Allowance and the natural log of median property price in each IG.

As the disease data were counts, I used Poisson log-linear models for the analysis. In the model, the random effects were modelled using conditional autoregressive (CAR) priors to account for residual spatio-temporal autocorrelation in the disease data after the known covariates have been accounted for.

The models were fitted under a Bayesian framework with the McMC being computed using one of two basic algorithms, the Gibbs sampling algorithm and the Metropolis-Hastings (M-H) algorithm. More details about this thesis are given as follows.

7.1 Initial impression of the air pollution health effects

The initial impression of the air pollution health effects was conducted in chapter 3, which used a benchmark method to investigate the health effects of NO_2 and PM_{10} individually. In this benchmark method, the pollution data from the monitoring network were not used, instead, only the modelled grid pollution data (DEFRA) have been used. The DEFRA data were converted into the IG scales on which the disease data were collected, by computing the spatial mean (or maximum) concentrations over the modelled grid data lying within each small area.

Note that the spatial mean is almost exclusively used as the aggregation function to transfer the DEFRA pollution data into a single metric for each IG (e.g. Maheswaran et al. [84] and Lee et al. [76]). In this thesis, I investigated both spatial mean and maximum metrics as it may be that peak concentrations are more suitable to be used to represent population exposure.

The results indicated that significant excess relative risks of respiratory hospital admissions were associated with long-term exposures to NO_2 or PM_{10} across IGs in mainland Scotland. Specifically, with a $6.84\mu gm^{-3}$ increase of peak NO_2 concentration, the hospital admissions related to respiratory disease in each IG will increase about 2.6% (ranges from 2.1% to 3.4%), while a significant increase was not found using the spatial mean NO_2. With a $1.872\mu gm^{-3}$ increase of mean PM_{10} concentration in the air, the hospital admissions related to respiratory disease in each IG will increase about 5.4% (ranges from 5.1% to 5.6%), while it is 3.7% (ranges from 3.5% to 4.0%) using the spatial maximum PM_{10} metric. These results are broadly consistent with those from other recent studies.

One obvious shortcoming for this benchmark method is that only the DEFRA data had been used to estimate exposure, which ignored the measured observations, and the latter are known to be more reliable.

7.2 Improved air pollution predictions - single-pollutant model

In chapter 4 I proposed a single-pollutant model, which was a novel statistical fusion model and enabled the use of both DEFRA and monitoring measurements to make point-level predictions of pollution across my study region, and finally aggregate these point-level predictions to the areal level to get the exposure for each IG.

The single-pollutant model was a spatio-temporal model which allowed for temporal autocorrelation in the model parameters in adjacent years. Conversely, I did not assume the measured concentrations were spatially autocorrelated after accounting for the covariate effects even though they are spatial data, because the exploratory analysis using geostatistical models provided little evidence for the presence of such autocorrelation after accounting for covariate effects (including DEFRA data). In order to assess the validity of this modelling approach, I compared my proposed model against using DEFRA data in isolation and the spatio-temporal pollution model (**SGH**) proposed by Sahu et al. [112] which did allow for residual spatial autocorrelation.

I measured the predictive performance using a 10-fold cross validation approach, and then quantified model performance by computing the prediction bias, root mean square prediction error (RMSPE) and the coverage probabilities of the 95% prediction intervals. The results showed that the single-pollutant model gave negligible bias, and model **SGH** had an RMSPE that was around 24% higher than that from my proposed single-pollutant model, despite all models having the same covariates. This is because the spatial random effects in Model **SGH** were competing with the covariates to explain the variation in the response, resulting in attenuation in the estimated covariate effects. The results also showed that using the DEFRA concentrations in isolation results in poorer spatial prediction than using both sources of data, with a RMSPE of 0.86 compared with 0.31 for the models proposed here. Finally, it is also shown that the DEFRA concentrations were an important covariate in the air pollution model as they resulting a reduced RMSPE.

7.3 Improved air pollution predictions - multi-pollutant model

In chapter 5, the multi-pollutant model which extended the single-pollutant model in chapter 4 was proposed to predict multiple pollutant concentrations. It allowed the correlation among pollutants to help improve the prediction of one pollutant by borrowing strength from the others.

The multi-pollutant model was also a spatio-temporal model which allowed the regression parameters to be temporally autocorrelated. Similar to the single-pollutant model, the spatial correlation among the observations was not considered, as the pollution observations across mainland Scotland did not have any residual spatial correlation after accounting for covariate effects.

The performance of this multi-pollutant model was good since the simulation study showed that the model parameters were estimated without bias, the RMSE of each parameter was low, and the coverage of each parameter was quite close to its nominal 95% level. Furthermore, the validation study showed that the multi-pollutant model outperformed the single pollutant model proposed in chapter 4 in terms of the RMSE, with improvements of 14% and 16% for NO_2 and PM_{10}, respectively.

7.4 Single pollutant health effects

I investigated the single pollutant health effects by fitting a single disease model (4.21) using each pollutant individually. Three types of pollution data had been considered in this study, DEFRA data, predictions from the single pollutant model and predictions from the multi-pollutant model. It is believed that the quality of the pollution predictions from the multi-pollutants model was better than the other two, as the multi-pollutant model used the correlation among pollutants to improve prediction. In addition, the quality of the pollution predictions from the single pollutant model were also better than the DEFRA data, as the former predicted the pollution using both the monitoring measurements and the DEFRA data. Therefore, it is of interest to compare the estimated relative risks while using different types of pollution data. The results from the DEFRA data were shown in chapter 3 which were also reviewed in section 7.1. The

results from fitting to the predictions from the single pollutant model and the multi-pollutant model were shown in chapter 4 and 6, respectively.

For NO_2, the statistical significance of the estimated relative risk was consistent for the three types of pollution data. That is the peak NO_2 concentration in the air was significantly associated with the respiratory hospital admissions in each IG, while the spatial mean NO_2 concentration was not. The similarity of the results across pollutant predictions validates the use of the DEFRA NO_2 data in my study even though they are known to be biased. The results from using the predictions from the multi-pollutant model suggested that a 6.84 μgm^{-3} increase in peak NO_2 exposure was associated with 3% (1.3%, 4.8%) higher respiratory disease hospital admissions in Scotland, whereas no relationship was observed when the spatial mean NO_2 was used.

For PM_{10}, the statistical significance of the estimated relative risk was also consistent for both DEFRA data and the predictions from the multi-pollutants model. That is both the peak and mean PM_{10} concentrations in the air were significantly associated with the respiratory hospital admissions in each IG. The use of the DEFRA PM_{10} data in my study was also validated. The results from using the predictions from the multi-pollutant model suggested that a 1.872 μgm^{-3} increase in peak PM_{10} exposure was associated with 5.6% (4.6%, 6.7%) higher respiratory disease hospital admissions in Scotland, while it was 1.7% (0.5%, 2.9%) higher for a 1.872 μgm^{-3} increase in mean PM_{10}.

7.5 Multiple pollutants health effects

I investigated the impact of the exposure to both NO_2 and PM_{10} simultaneously in chapter 6. As NO_2 and PM_{10} were highly correlated to each other in my study, I first regressed NO_2 (or PM_{10}) on PM_{10} (or NO_2) and used its residuals in the disease model as well as PM_{10} (or NO_2), thus investigating the health effects of exposure to both pollutants simultaneously. These residuals were interpreted as the remaining signal from NO_2 (or PM_{10}) which could not be explained by PM_{10} (or NO_2), however, it was difficult to interpret their effects in the multi-pollutant disease model.

The output from the multiple pollutant disease model confirmed the adverse effects of both NO_2 and PM_{10} from the single pollutant studies. It showed that a 6.84 μgm^{-3}

increase in peak NO_2 exposure was associated with 3.4% (2.1%, 4.6%) higher respiratory disease hospital admissions in Scotland, whereas no relationship was observed when the spatial mean NO_2 was used, and a 1.872 μgm^{-3} increase in peak PM_{10} exposure was associated with 3.3% (2.4%, 4.3%) higher respiratory disease hospital admissions in Scotland, whereas it was 1.4% (0.3%, 2.4%) for a 1.872 μgm^{-3} increase in mean PM_{10} exposure.

The results also indicated that the remaining signal from both spatial mean and maximum PM_{10} which could not be explained by NO_2, were not helpful to explain the residuals from the disease data after accounting for NO_2 and other covariates. Similarly, the remaining signal from spatial mean NO_2 which could not be explained by spatial mean PM_{10} was not significant in the disease model, given the PM_{10} pollutant and other covariates. However, by using the spatial maximum metric, the remaining signal from NO_2 was significant in the disease model. This is likely because the residual variable in this model contained a non-ignorable signal which could help to explain the disease data after accounting for the PM_{10} and other covariates.

7.6 Dealing with exposure uncertainty

In this thesis, I developed and compared four approaches to incorporate the exposure uncertainty outlined in chapter 6. The first one was to fit the multi-pollutant disease model 100 times, with a different posterior predictive sample for pollution each time. The second approach was to incorporate the uncertainty of the exposure into the muti-pollutant disease model by allowing the exposure to be sampled from the $M = 100$ sets of predictions for each iteration of the McMC algorithm. The third approach was to add a classical measurement error model as an extra level in the multi-pollutant disease model, while the last approach was similar to the third approach except that the variance of exposure was allowed to be varied across IGs and time periods which matched the real data set better.

The results suggested that both approaches 1 and 2 appear to perform poorly as the significant relative risks observed previously have disappeared, due to the variation of the exposure estimates within IGs being comparable to that across IGs. In contrast, both approaches 3 and 4 maintain the significant relative risks while allowing the exposure

errors. The estimates of the relative risk of exposure from both approaches 3 and 4 were consistent, that spatial maximum NO_2, spatial mean and maximum PM_{10} were all significantly associated with the respiratory hospital admissions. In addition, the uncertainty of the relative risk of X_{1kt} is reduced compared to the multi-pollutant disease model which doesn't consider the uncertainty of exposure, indicating that the health data are informative for the update of pollution as the pollution data in approach 4 were allowed to be updated depending on the health data. Note that both approaches 3 and 4 are computationally expensive due to the requirement of updating a large number of parameters (as the exposure for each IG and each period is treated as an unknown parameter) in each iteration of the McMC algorithm.

7.7 Key themes

This thesis contributes to epidemiological studies by improving our understanding in a few aspects. The pollution predictions are improved by fusing multiple sources of pollution data, with the example in this thesis that the predicted NO_2 concentrations from a fusion model of both measured and DEFRA data are better than the standard DEFRA data. The prediction of one pollutant can be improved by borrowing the strength from the others, which is shown by the validation study of the proposed multi-pollutant model in chapter 5. Although the pollution concentrations can be improved by fusing different sources of pollution data, the health effects are largely consistent between using these fused pollution concentrations and the standard DEFRA data. This validates the use of DEFRA data in the study of air pollution health effects.

Exposure uncertainty is a key aspect in the study of air pollution health effects as the exposures are only estimates and subject to uncertainty which needs to be accounted for. This thesis develops several new methods to propagate the exposure uncertainty into the health effects model. However, the exposure uncertainty in epidemiological studies is an under researched topic which needs more work, such as the investigation of its effects via simulation studies.

We also learn that there are independent health effects for different pollutants, which is shown in this study, since the remaining signal from NO_2 which cannot be explained by

PM_{10} is significantly associated with the respiratory hospital admissions after accounting for PM_{10} and other covariates effects.

7.8 Discussions and future work

There are several limitations and possible extensions to the statistical analysis of air pollution health effects that has been carried out within this thesis. The nature of possible future work involves not only direct extensions of the analysis of the problems presented, but also could involve additional statistical challenges.

There is a limitation of the design of the monitoring network in my study, where the monitor locations are highly clustered in urban area while no monitors exist in large parts of the study region (see Figure 4.1). Therefore, in chapter 4, the predictive performance for those exposure models cannot be assessed uniformly across Scotland, which lead to the uncertainty of the prediction performance at rural areas where no monitors exist. However, as it is known that the NO_2 concentrations are low in rural regions where the traffic is few, so the level of uncertainty should be low and the DEFRA concentrations should be able to pick up the low background levels.

Another potential limitation lies in the socio-economic deprivation confounders used in this thesis, which are the percentage of people in receipt of job seekers allowance and the natural log of the median property price. Socio-economic deprivation is difficult to measure because it includes various aspects such as income, education and housing, and these variables are potentially highly correlated. For example, the Index of Multiple Deprivation is an alternative way of measuring socio-economic deprivation. Recent research by Pannullo et al. [98] reported that the different measures of deprivation can result in a variety of pollution-health effects. Therefore, it will be worth extending my current work by investigating how different measures of deprivation affect the pollution-health effects.

Through the thesis, in disease models the conditional autoregressive (CAR) models were specified as a prior distribution for a set of random effects, and the spatial correlation structure induced by these models was determined by geographical adjacency, which means that two areas have correlated random effects if they share a common border. A challenge which is also a limitation lying in this assumption is that two geographically

adjacent IGs might have very different risk profiles because of different deprivation levels or maybe living habits. For example, Lee and Mitchell [78] identified such risk boundaries in their study in Greater Glasgow. In this case, the random effects for these IGs could be uncorrelated rather than correlated which is assumed by the neighbouring effects. Therefore, a potential avenue of future work could be the investigation of the effects using different definitions of neighbours (e.g. two IGs are defined to be neighbours if their central points are within a fixed distance, or if one area is one of the h closest areas to another area in terms of distance).

In this study, the spatial mean or maximum concentrations in each IG have been used to represent the exposure of the people living there. However, this estimated exposure could be very different to the real exposure. For example, the air pollution concentrations are very different from indoors to outdoors and the time for each individual to be outdoors depends mainly on their jobs. Another example is that some people work far away from where they live and in this case, the real exposure for these people should be estimated as a combination of where they work and where they live. Therefore, a further avenue of future work could be new methodologies to improve exposure estimation.

Note that the respiratory disease is not the only health consequence of long-term exposure to air pollution. In addition, because poor respiratory health can contribute to other serious diseases, the true health burden to which air pollution may contribute is likely to be far larger than that estimated. Therefore, an interesting question to answer in air pollution health studies is: What is the health burden of air pollution? Such studies would require a multivariate disease model which considers the pollution impacts on different types of diseases rather than only one specific disease type.

My study showed the consistency between the estimated health effects while using two types of pollution data, the DEFRA concentrations alone and both the measured and DEFRA concentrations. This consistency was observed when considering both the spatial mean and maximum as the aggregation functions, and suggested that the DEFRA concentrations appeared reliable to use in health effect studies despite being biased. Therefore, it could be of interest to examine whether this result is widely true for other study regions and pollutants, or whether it is not always so consistent. This reliability of the DEFRA data is a key question, because its widespread availability makes it a

popular choice for health effect studies, especially when the measured data are spatially sparse.

It is of interest to extend the study to a wider or another region, e.g. the whole UK, so as to know whether the adverse effects of NO_2 and PM_{10} are general or just local to Scotland. Another extension could be the investigation of air pollution health effects based on a finer temporal scale, such as monthly, so that the long-term effects of air pollution on health can be investigated with different lags between the occurence of disease and the exposure. Such studies would be helpful to understand how the relationship between exposure and human health changes by time.

Finally, the residuals from the temporally-varying linear model (6.2) are interpreted as the remaining signal from ($\boldsymbol{X}_2$) which can not be explained by ($\boldsymbol{X}_1$), and the residual coefficient in the multi-pollutant disease model can be interpreted as the impact of the variation in one pollutant which can not be explained by another pollutant on the risk. Therefore, $\boldsymbol{X}_1$ (main pollutant in the multi-pollutant disease model) should represent the main pollutant we want to investigate. Theoretically, we can use the same methodology to explore simultaneously as many pollutants as we want. However, a challenge for such studies could be the availability of pollution data. In practice, very few pollutants have dense networks of monitors to cover a big area (e.g. in my study, the other pollutants besides NO_2 and PM_{10} are sparsely measured across Scotland), so it is more likely to investigate such multiple pollutant health effects based on a relative small region. The proposed multi-pollutant disease model in chapter 6 can be extended to deal with more than two pollutants. I use three pollutants ($\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$) as an example to explain how to extend the multi-pollutant disease model to handle more than two pollutants. Firstly, regress $\boldsymbol{X}_2$ against $\boldsymbol{X}_1$ to get the residuals $\hat{\boldsymbol{\epsilon}}_{21}$ which are uncorrelated to $\boldsymbol{X}_1$. Then, regress $\boldsymbol{X}_3$ against both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to get the residuals $\hat{\boldsymbol{\epsilon}}_{321}$ which are uncorrelated to both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ ($\hat{\boldsymbol{\epsilon}}_{321} \cdot \boldsymbol{X}_1 = \hat{\boldsymbol{\epsilon}}_{321} \cdot \boldsymbol{X}_2 = 0$). As $\hat{\boldsymbol{\epsilon}}_{21} = \boldsymbol{X}_2 - \beta \boldsymbol{X}_1$ (β is a constant),

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_{321} \cdot \hat{\boldsymbol{\epsilon}}_{21} &= \hat{\boldsymbol{\epsilon}}_{321} \cdot (\boldsymbol{X}_2 - \beta \boldsymbol{X}_1) \\ &= \hat{\boldsymbol{\epsilon}}_{321} \cdot \boldsymbol{X}_2 - \hat{\boldsymbol{\epsilon}}_{321} \cdot \beta \boldsymbol{X}_1 \\ &= 0 - 0 \\ &= 0 \end{aligned}$$

Therefore, $\hat{\boldsymbol{\epsilon}}_{321}$ is uncorrelated to $\hat{\boldsymbol{\epsilon}}_{21}$, and $\boldsymbol{X}_1$, $\hat{\boldsymbol{\epsilon}}_{21}$, $\hat{\boldsymbol{\epsilon}}_{321}$ can be put into a single disease model without causing multicollinearity. Again, a challenge for using this multi-pollutant disease model to investigate the health effects of exposure to multiple pollutants is the interpretation of the results, e.g. how to interpret the effects of $\hat{\boldsymbol{\epsilon}}_{321}$. Note that the multi-pollutant model proposed in chapter 5 will be used to predict pollution prior to fitting the multi-pollutant disease model, where the computational challenge likely occurs. In my study, the spatial correlation between the monitoring data are very weak, therefore an approximate method used to update the missing observation at each site is adopted, which assumed the distribution of a missing value at a monitoring site only depended on another pollutant at the same site rather than all the observations across all sites. Under this approximation, it took about 3 hours to predict the 1km gridded concentrations for both NO_2 and PM_{10} across mainland Scotland. If more pollutants are considered and the spatial correlation between monitoring data is no longer ignorable, the computational burden could increase dramatically. In addition, more pollutants also means the increase of computational burden in the implementation of the multi-pollutant disease model, especially those models proposed in approach 3 and 4 where the exposure estimates are updated in each iteration of the McMC algorithm.

Bibliography

[1] AEA (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, benezene, CO and ozone. http://uk-air.defra.gov.uk/assets/documents/reports/cat09/1204301513_AQD2010mapsrep_master_v0.pdf.

[2] Allodji, R. S., K. Leuraud, A. C. M. Thiébaut, S. Henry, D. Laurier, and J. Bénichou (2012). Impact of measurement error in radon exposure on the estimated excess relative risk of lung cancer death in a simulated study based on the French Uranium Miners' Cohort. *Radiation and Environmental Biophysics 51*(2), 151 – 163.

[3] Arif, A. and S. Shah (2007). Association between personal exposure to volatile organic compounds and asthma among US adult population. *International Archives of Occupational and Environmental Health 80*(8), 711 – 719.

[4] Armstrong, B. G. (1990). The effects of measurement errors on relatwe risk regressions. *American Journal of Epidemiology 132*(6), 1176 – 1184.

[5] Barceló, M. A., M. Saez, and C. Saurina (2009). Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. *Science of The Total Environment 407*(21), 5501 – 5523.

[6] Basu, R., T. J. Woodruff, J. D. Parker, L. Saulnier, and K. C. Schoendorf (2004). Comparing exposure metrics in the relationship between PM2.5 and birth weight in California. *Journal of Exposure Analysis and Environmental Epidemiology 14*, 391 – 396.

[7] Beelen, R., O. Raaschou-Nielsen, M. Stafoggia, Z. J. Andersen, G. Weinmayr, B. Hoffmann, K. Wolf, E. Samoli, P. Fischer, M. Nieuwenhuijsen, P. Vineis, W. W. Xun, K. Katsouyanni, K. Dimakopoulou, A. Oudin, B. Forsberg, L. Modig, A. S.

Havulinna, T. Lanki, A. Turunen, B. Oftedal, W. Nystad, P. Nafstad, U. D. Faire, N. L. Pedersen, C.-G. stenson, L. Fratiglioni, J. Penell, M. Korek, G. Pershagen, K. T. Eriksen, K. Overvad, T. Ellermann, M. Eeftens, P. H. Peeters, K. Meliefste, M. Wang, B. B. de Mesquita, D. Sugiri, U. Krmer, J. Heinrich, K. de Hoogh, T. Key, A. Peters, R. Hampel, H. Concin, G. Nagel, A. Ineichen, E. Schaffner, N. Probst-Hensch, N. Knzli, C. Schindler, T. Schikowski, M. Adam, H. Phuleria, A. Vilier, F. Clavel-Chapelon, C. Declercq, S. Grioni, V. Krogh, M.-Y. Tsai, F. Ricceri, C. Sacerdote, C. Galassi, E. Migliore, A. Ranzi, G. Cesaroni, C. Badaloni, F. Forastiere, I. Tamayo, P. Amiano, M. Dorronsoro, M. Katsoulis, A. Trichopoulou, B. Brunekreef, and G. Hoek (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre {ESCAPE} project. *The Lancet 383*(9919), 785 – 795.

[8] Belanger, K., J. F. Gent, E. W. Triche, M. B. Bracken, and B. P. Leaderer (2006). Association of indoor nitrogen dioxide exposure with respiratory symptoms in children with asthma. *American journal of respiratory and critical care medicine 173*(3), 297 – 303.

[9] Bell, M. and D. Davis (2001). Reassessment of the Lethal London Fog of 1952: Novel Indicators of Acute and Chronic Consequences of Acute Exposure to Air Pollution. *Environmental Health Perspectives 109*(3), 389 – 394.

[10] Bennett, J., M. P. Little, and S. Richardson (2004). Flexible dose-response models for Japanese atomic bomb survivor data: Bayesian estimation and prediction of cancer risk. *Radiation and Environmental Biophysics 43*(4), 233 – 245.

[11] Berrocal, V., A. Gelfand, and D. Holland (2010a). A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics 15*(2), 176 – 197.

[12] Berrocal, V., A. Gelfand, D. Holland, J. Burke, and M. L. Miranda (2011). On the use of a PM2.5 exposure simulator to explain birthweight. *Environmetrics 22*(4), 553 – 571.

[13] Berrocal, V. J., A. E. Gelfand, and D. M. Holland (2010b, 12). A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics 4*(4), 1942 – 1975.

[14] Berrocal, V. J., A. E. Gelfand, and D. M. Holland (2012). Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. *Biometrics 68*(3), 837 – 848.

[15] Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics 43*, 1 – 59.

[16] Blair, A., P. Stewart, J. H. Lubin, and F. Forastiere (2007). Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *American Journal of Industrial Medicine 50*(3), 199 – 207.

[17] Bobb, J. F., L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*.

[18] Brown, T. P., L. Rushton, M. A. Mugglestone, and D. F. Meechan (2003). Health effects of a sulphur dioxide air pollution episode. *Journal of Public Health 25*(4), 369 – 371.

[19] Brunekreef, B., D. Noy, and P. Clausing (1987). Variability of exposure measurements in environmental epidemiology. *American Journal of Epidemiology 125*(5), 892 – 898.

[20] Bruno, F. and D. Cocchi (2002). A unified strategy for building simple air quality indices. *Environmetrics 13*(3), 243 – 261.

[21] Bruno, F., D. Cocchi, F. Greco, and E. Scardovi (2013). Spatial reconstruction of rainfall fields from rain gauge and radar data. *Stochastic Environmental Research and Risk Assessment*, 1 – 11.

[22] Carder, M., R. McNamee, I. Beverland, R. Elton, M. V. Tongeren, G. R. Cohen, J. Boyd, W. MacNee, and R. M. Agius (2008). Interacting effects of particulate pollution and cold temperature on cardiorespiratory mortality in Scotland. *Occupational and environmental medicine 65*, 197 – 204.

[23] Carey, I., R. Atkinson, A. Kent, T. van Staa, D. Cook, and H. Anderson (2013). Mortality associations with long-term exposure to outdoor air pollution in a national English cohort. *Am J Respir Crit Care Med 187*(11), 1226 – 33.

[24] Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

[25] Cesaroni, G., C. Badaloni, C. Gariazzo, M. Stafoggia, R. Sozzi, M. Davoli, and F. Forastiere (2013). Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environ. Health Perspect. 121*, 324 – 331.

[26] Chauhan, A., M. Krishna, A. Frew, and S. Holgate (1998). Exposure to nitrogen dioxide (NO2) and respiratory disease risk. *Reviews on Environmental Health 13*, 73 – 90.

[27] Chen, R., W. Huang, C. Wong, Z. Wang, T. Thach, B. Chen, and H. Kan (2012). Short-term exposure to sulfur dioxide and daily mortality in 17 Chinese cities: The China air pollution and health effects study (capes). *Environmental Research 118*, 101 – 106.

[28] Coker, E., S. Liverani, J. K. Ghosh, M. Jerrett, B. Beckerman, A. Li, B. Ritz, and J. Molitor (2016). Multi-pollutant exposure profiles associated with term low birth weight in los angeles county. *Environment International 91*, 1 – 13.

[29] Committee on the Medical Effects of Air Pollutants (2010). *The Mortality Effects of Long-Term Exposure to Particulate Air Pollution in the United Kingdom*. Crown.

[30] Cowles, M. and B. P. Carlin (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association 91*(434), 883 – 904.

[31] Cressie, N. (1993). *Statistics for Spatial Data* (revised ed.). New York: Wiley.

[32] Department for Environment and Food And Rural Affairs (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland. (The Stationery Office.

[33] Department for Environment and Food And Rural Affairs (2015). Draft plans to improve air quality in the UK - Tackling nitrogen dioxide in our towns and cities. Crown.

[34] Desqueyroux, H., J.-C. Pujet, M. Prosper, F. Squinazi, and I. Momas (2002). Short-Term Effects of Low-Level Air Pollution on Respiratory Health of Adults Suffering from Moderate to Severe Asthma. *Environmental Research 89*(1), 29 – 37.

[35] Diepgen, T. and V. Mahler (2002). The epidemiology of skin cancer. *British Journal of Dermatology 146*, 1 – 6.

[36] Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics*. Springer Series in Statistics, Springer.

[37] Dockery, D. W., C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. J. Ferris, and F. E. Speizer (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine 329*(24), 1753 – 1759.

[38] Dominici, F., M. Daniels, S. L. Zeger, and J. M. Samet (2002). Air Pollution and Mortality. *Journal of the American Statistical Association 97*(457), 100 – 111.

[39] Dominici, F., S. L. Zeger, and J. M. Samet (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics 1*(2), 157–175.

[40] Ehrlich, R., J. Findlay, J. Fenters, and D. Gardner (1977). Health effects of short-term inhalation of nitrogen dioxide and ozone mixtures. *Environmental Research 14*, 223 – 231.

[41] Elliott, P., G. Shaddick, J. Wakefield, C. de Hoogh, and D. Briggs (2007). Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax 62*(12), 1088 – 1094.

[42] Fuentes, M. and A. E. Raftery (2005). Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models. *Biometrics 61*(1), 36 – 45.

[43] Gelfand, A. E. and A. F. M. Smith (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association 85*(410), 398 – 409.

[44] Gelfand, A. E., L. Zhu, and B. P. Carlin (2001). On the change of support problem for spatio-temporal data. *Biostatistics 2*(1), 31 – 45.

[45] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 1 – 19.

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

[47] Gilmour, P. S., A. Ziesenis, E. Morrison, M. A. Vickers, E. M. Drost, I. Ford, E. Karg, C. Mossa, A. Schroeppel, G. A. Ferron, J. Heyder, M. Greaves, W. MacNee, and K. Donaldson (2004). Pulmonary and systemic effects of short-term inhalation exposure to ultrafine carbon black particles. *Toxicology and Applied Pharmacology 195*(1), 35 – 44.

[48] Gladen, B. and W. J. Rogan (1979). Misclassification and the design of environmental studies. *American Journal of Epidemiology 109*(5), 607 – 616.

[49] Goldberg, M. S. and R. T. Burnett (2005). A New Longitudinal Design for Identifying Subgroups of The Population who are Susceptible to the Short-term Effects of Ambient Air Pollution. *Journal of Toxicology and Environmental Health, Part A 68*(13 -14), 1111 – 1125.

[50] Gotway, C. A. and L. J. Young (2002). Combining Incompatible Spatial Data. *American Statistical Association 97*(458), 632 – 648.

[51] Government, T. S. (2010). *Scottish Government Urban/Rural Classification 2009 - 2010*. Scottish Government.

[52] Greven, S., F. Dominici, and S. Zeger (2011). An Approach to the Estimation of Chronic Air Pollution Effects Using Spatio-Temporal Information. *Journal of the American Statistical Association 106*(494), 396 – 406.

[53] Haining, R., G. Li, R. Maheswaran, M. Blangiardo, J. Law, N. Best, and S. Richardson (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*(2 - 3), 123 – 131.

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

[55] Health and Safety Executive (2014). *Ozone: Health hazards and control measures EH38 (Third edition)*. Crown.

[56] Heid, I., H. Kuchenhoff, J. Miles, L. Kreienbrock, and H. Wichmann (2004). Two dimensions of measurement error: Classical and Berkson errorin residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology 14*, 365 – 377.

[57] Heidelberger, P. and P. D. Welch (1981, April). A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations. *Commun. ACM 24*(4), 233 – 245.

[58] Heidelberger, P. and P. D. Welch (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research 31*(6), 1109 – 1144.

[59] Hoek, G., B. Brunekreef, S. Goldbohm, P. Fischer, and P. van den Brandt (2002). Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *The Lancet 360*(9341), 1203 – 1209.

[60] Hoek, G., P. Fischer, P. van den Brandt, S. Goldbohm, and B. Brunekreef (2001). Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. *Journal of Exposure Analysis and Environmental Epidemiology* (11), 459 – 469.

[61] Hong, Y. C., J. H. Leem, E. H. Ha, and D. C. Christiani (1999). PM10 exposure, gaseous pollutants, and daily mortality in Inchon, South Korea. *Environmental Health Perspectives 107*(11), 873 – 878.

[62] Janes, H., F. Dominici, and S. Zeger (2007). Trends in Air Pollution and Mortality: An Approach to the Assessment of Unmeasured Confounding. *Epidemiology 18*(4), 416 – 423.

[63] Jerrett, M., M. Buzzelli, R. T. Burnett, and P. F. DeLuca (2005). Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Social Science & Medicine 60*(12), 2845 – 2863.

[64] Journel, A. G. and C. J. C. J. Huijbregts (2003). *Mining geostatistics*. New York : The Blackburn Press.

[65] Katsouyanni, K., G. Touloumi, E. Samoli, A. Gryparis, A. Le Tertre, Y. Monopolis, G. Rossi, D. Zmirou, F. Ballester, A. Boumghar, H. Anderson, B. Wojtyniak, A. Paldy, R. Braunstein, J. Pekkanen, C. Schindler, and J. Schwartz (2001). Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 project. *Epidemiology 12*(5), 521 – 531.

[66] Kattan, M., P. J. Gergen, P. Eggleston, C. M. Visness, and H. E. Mitchell (2007). Health effects of indoor nitrogen dioxide and passive smoking on urban asthmatic children. *The Journal of allergy and clinical immunology 120*(3), 618 – 624.

[67] Kim, J. Y., R. T. Burnett, L. Neas, G. D. Thurston, J. Schwartz, P. E. Tolbert, B. Brunekreef, M. S. Goldberg, and I. Romieu (2007). Panel discussion review: session two - interpretation of observed associations between multiple ambient air pollutants and health effects in epidemiologic analyses. *Journal of Exposure Science and Environmental Epidemiology 17*, 83–89.

[68] Kioumourtzoglou, M. A., D. Spiegelman, A. A. Szpiro, L. Sheppard, J. D. Kaufman, J. D. Yanosky, R. Williams, F. Laden, B. Hong, and H. Suh (2014). Exposure measurement error in PM2.5 health effects studies: A pooled analysis of eight personal exposure validation studies. *Environmental Health 13*(1), 1 – 11.

[69] Kontos, A., S. Fassois, and M. Deli (1999). Short-Term Effects of Air Pollution on Childhood Respiratory Illness in Piraeus, Greece, 1987 - 1992: Nonparametric Stochastic Dynamic Analysis. *Environmental Research 81*(4), 275 – 296.

[70] Kumar, R. and A. Joseph (2006). Air Pollution Concentrations of PM2.5, PM10 and NO2 at Ambient and Kerbsite and Their Correlation in Metro City Mumbai. *Environmental Monitoring and Assessment 119*(1 - 3), 191 – 199.

[71] Laden, F., J. Schwartz, F. E. Speizer, and D. W. Dockery (2006). Reduction in Fine Particulate Air Pollution and Mortality. *American Journal of Respiratory and Critical Care Medicine 173*(6), 667 – 672.

[72] Lawson, A., J. Choi, B. Cai, M. Hossain, R. Kirby, and J. Liu (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*(3), 417 – 441.

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

[74] Lee, D. (2012). Using spline models to estimate the varying health risks from air pollution across Scotland. *Statistics in Medicine 31*(27), 3366 – 3378.

[75] Lee, D. (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. *Journal of Statistical Software 55*(13), 1 – 24.

[76] Lee, D., C. Ferguson, and R. Mitchell (2009). Air pollution and health in Scotland: a multicity study. *Biostatistics 10*(3), 409 – 423.

[77] Lee, D., C. Ferguson, and E. M. Scott (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(1), 109 – 126.

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

[79] Lee, H., Y. Honda, M. Hashizume, Y. L. Guo, C.-F. Wu, H. Kan, K. Jung, Y.-H. Lim, S. Yi, and H. Kim (2015). Short-term exposure to fine and coarse particles and mortality: A multicity time-series study in East Asia. *Environmental Pollution 207*, 43 – 51.

[80] Leroux, B., X. Lei, and N. Breslow (1999). *Estimation of disease rates in small areas: A new mixed model for spatial dependence*, Chapter Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds), pp. 135–178. Springer-Verlag, New York.

[81] Li, Y., A. Guolo, F. O. Hoffman, and R. J. Carroll (2007). Shared Uncertainty in Measurement Error Problems, with Application to Nevada Test Site Fallout Data. *Biometrics 63*(4), 1226 – 1236.

[82] MacLehose, R. F., D. B. Dunson, A. H. Herring, and J. A. Hoppin (2007). Bayesian Methods for Highly Correlated Exposure Data. *Epidemiology 18*(2), 199–207.

[83] Macnab, Y. (2003). Hierarchical Bayesian modelling of spatially correlated health service outcome and utilization rates. *Biometrics 59*, 305 – 316.

[84] Maheswaran, R., R. Haining, T. Pearson, J. Law, P. Brindley, and N. Best (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.

[85] Maheswaran, R., R. P. Haining, P. Brindley, J. Law, T. Pearson, P. R. Fryers, S. Wise, and M. J. Campbell (2005). Outdoor Air Pollution and Stroke in Sheffield, United Kingdom: A Small-Area Level Geographical Study. *Stroke 36*(2), 239 – 243.

[86] Matérn, B. (1960). *Spatial Variation*. Springer-Verlag New York.

[87] McMillan, N. J., D. M. Holland, M. Morara, and J. Feng (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics 21*(1), 48 – 65.

[88] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Chemical Physics 21*, 1087 – 1091.

[89] Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

[90] Molitor, J., M. Jerrett, C.-C. Chang, N.-T. Molitor, J. Gauderman, K. Berhane, R. McConnell, F. Lurmann, J. Wu, A. Winer, and D. Thomas (2007). Assessing Uncertainty in Spatial Exposure Models for Air Pollution Health Effects Assessment. *Environmental Health Perspectives 115*(8), 1147 – 1153.

[91] Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*, 17 – 23.

[92] Nafstad, P., L. L. Hheim, T. Wisloff, F. Gram, B. Oftedal, I. Holme, I. Hjermann, and P. Leren (2004). Urban air pollution and mortality in a cohort of norwegian men. *Environ Health Perspect 112*(5), 610 – 5.

[93] Neukirch, F., C. Sgala, Y. L. Moullec, M. Korobaeff, and M. Aubier (1998). Short-Term Effects of Low-Level Winter Pollution on Respiratory Health of Asthmatic Adults. *Archives of Environmental Health: An International Journal 53*(5), 320 – 328.

[94] Oakes, M., L. Baxter, and T. C. Long (2014). Evaluating the application of multipollutant exposure metrics in air pollution health studies. *Environment International 69*(0), 90 – 99.

[95] Oftedal, B., P. Nafstad, P. Magnus, S. Bjørkly, and A. Skrondal (2003). Traffic related air pollution and acute hospital admission for respiratory diseases in Drammen, Norway 1995–2000. *European Journal of Epidemiology 18*(7), 671 – 676.

[96] Pachon, J. E., S. Balachandran, Y. Hu, J. A. Mulholland, L. A. Darrow, J. A. Sarnat, P. E. Tolbert, and A. G. Russell (2012). Development of outcome-based, multipollutant mobile source indicators. *Journal of the Air & Waste Management Association 62*(4), 431 – 442.

[97] Pannullo, F., D. Lee, E. Waclawski, and A. H. Leyland (2015). Improving spatial nitrogen dioxide prediction using diffusion tubes: A case study in West Central Scotland. *Atmospheric Environment 118*, 227 – 235.

[98] Pannullo, F., D. Lee, E. Waclawski, and A. H. Leyland (2016). How robust are the estimated effects of air pollution on health? Accounting for model uncertainty using bayesian model averaging. *Spatial and Spatio-temporal Epidemiology*, –.

[99] Papathomas, M., J. Molitor, S. Richardson, E. Riboli, and P. Vineis (2011). Examining the joint effect of multiple risk factors using exposure risk profiles: Lung cancer in nonsmokers. *Environmental Health Perspectives 119*(1), 84 – 91.

[100] Pickles, J. (1982). Air pollution estimation error and what it does to epidemiological analysis. *Atmospheric Environment (1967) 16*(9), 2241 – 2245.

[101] Pirani, M., N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller (2015). Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International 79*, 56 – 64.

[102] Pope, I. C., R. Burnett, M. Thun, and et al (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association 287*(9), 1132 – 1141.

[103] Powell, H. and D. Lee (2014). Modelling spatial variability in concentrations of single pollutants and composite air quality indicators in health effects studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 177*(3), 607 – 623.

[104] Prescott, G. J., G. R. Cohen, R. A. Elton, F. G. Fowkes, and R. M. Agius (1998). Urban air pollution and cardiopulmonary ill health: a 14.5 year time series study. *Occupational and environmental medicine 55*(10), 697 – 704.

[105] Qian, Z., J. Zhang, L. R. Korn, F. Wei, and R. S. Chapman (2004). Factor analysis of household factors: are they associated with respiratory conditions in Chinese children? *International Journal of Epidemiology 33*, 582 – 588.

[106] Raaschou-Nielsen, O., Z. J. Andersen, S. S. Jensen, M. Ketzel, M. Sørensen, J. Hansen, S. Loft, A. Tjønneland, and K. Overvad (2012). Traffic air pollution and mortality from cardiovascular disease and all causes: a Danish cohort study. *Environmental Health 11*(1), 1 – 12.

[107] Roberts, G. and A. Smith (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications 49*(2), 207 – 216.

[108] Roberts, G. O. and R. L. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika 83*(1), 95 – 110.

[109] Rushworth, A., D. Lee, and R. Mitchell (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.

[110] Ryan, T. P. (1997). *Modern regression methods*. John Wiley & Sons.

[111] Sacks, J. D., A. G. Rappold, J. Allen Davis Jr., D. B. Richardson, A. E. Waller, and T. J. Luben (2014). Influence of urbanicity and county characteristics on the association between ozone and asthma emergency department visits in North Carolina. *Environ Health Perspect 122*(5), 506 – 512.

[112] Sahu, S. K., A. E. Gelfand, and D. M. Holland (2007). High Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association 102*(480), 1221 – 1234.

[113] Sahu, S. K., A. E. Gelfand, and D. M. Holland (2010). Fusing point and areal level space-time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(1), 77 – 103.

[114] Schruben, L., H. Singh, and L. Tierney (1983). Optimal Tests for Initialization Bias in Simulation Output. *Oper. Res. 31*(6), 1167 – 1178.

[115] Schruben, L. W. (1982). Detecting initialization Bias in Simulation Output. *Operations Research 30*(3), 569 – 590.

[116] Seaton, A. and M. Dennekamp (2003). Hypothesis: Ill health associated with low concentrations of nitrogen dioxide-an effect of ultrafine particles? *Thorax 58*(12), 1012 – 1015.

[117] Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of The Royal Statistical Society Series C-applied Statistics 51*, 351 – 372.

[118] Shin, H., T. Ramsay, D. Krewski, and J. M. Zielinski (2005). The effect of censoring on cancer risk estimates based on the Canadian National Dose Registry of occupational radiation exposure. *Journal of Exposure Analysis and Environmental Epidemiology 15*, 398 – 406.

[119] Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 64*, 583 – 639.

[120] Spiegelman, D. (2010). Approaches to Uncertainty in Exposure Assessment in Environmental Epidemiology. *Annual review of public health 31*, 149 – 163.

[121] Stern, H. S. and N. Cressie (1999). *Inference for extremes in disease mapping*, Chapter Disease Mapping and Risk Assessment for Public Health.Lawson,A and Biggeri,D and Boehning,E and Lesaffre,E and Viel,J and Bertollini,R(eds). Willey.

[122] Stockfelt, L., E. M. Andersson, P. Molnr, A. Rosengren, L. Wilhelmsen, G. Sallsten, and L. Barregard (2015). Long term effects of residential NO(x) exposure on total and cause-specific mortality and incidence of myocardial infarction in a Swedish cohort. *Environmental Research 142*, 197 – 206.

[123] Stram, D. (2005). Designs for studies of personal exposure to air pollution and the impact of measurement error. *J Toxicol Environ Health A 68*, 1181 – 7.

[124] Tan, M. T., G.-L. Tian, and K. W. Ng (2010). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation.* Chapman & Hall/CRC.

[125] Tecer, L., O. Alagha, F. Karaca, G. Tuncel, and N. Eldes (2008). Particulate Matter (PM2.5, PM10-2.5, and PM10) and Children's Hospital Admissions for Asthma and Respiratory Diseases: A Bidirectional Case-Crossover Study. *Journal of Toxicology and Environmental Health, Part A 71*(8), 512 – 520.

[126] Thishan Dharshana, K. G. and N. Coowanitwong (2008). Ambient PM10 and respiratory illnesses in Colombo City, Sri Lanka. *Journal of Environmental Science and Health, Part A 43*(9), 1064 – 1070.

[127] Thomas, D., D. Stram, and J. Dwyer (1993). Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annu Rev Public Health 14*, 69 – 93.

[128] Thormod, H., D. Arne, D. Soren H.H., and M. Johan (1990). Ultraviolet-radiation and skin cancer. effect of an ozone layer depletion. *Photochemistry and Photobiology 51*(5), 579 – 582.

[129] To, T., S. Shen, E. G. Atenafu, J. Guan, S. McLimont, B. Stocks, and C. Licskai (2013). The Air Quality Health Index and Asthma Morbidity: A Population-Based Study. *Environmental Health Perspectives 121*, 46 – 52.

[130] Tolberta, P. E., M. Kleina, J. L. Peelb, S. E. Sarnata, and J. A. Sarnata (2007). Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta. *Journal of Exposure Science and Environmental Epidemiology 17*, 29 – 35.

[131] Townsend, N., P. Bhatnagar, E. Wilkins, K. Wickramasinghe, and M. Rayner (2015). *Cardiovascular disease statistics*. British Heart Foundation: London.

[132] Tunnicliffe, W., P. Burge, and J. Ayres (1994). Effect of domestic concentrations of nitrogen dioxide on airway responses to inhaled allergen in asthmatic patients. *The Lancet 344*(8939 - 8940), 1733 – 1736.

[133] US Environmental Protection Agency (2011). The Benefits and Costs of the Clean Air Act: 1990 to 2020. *Final Report of U.S. Environmental Protection Agency Office of Air and Radiation*, 5 – 10.

[134] Vinikoor-Imler, L. C., J. A. Davis, R. E. Meyer, and T. J. Luben (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*(10), 696 – 701.

[135] Vinikoor-Imler, L. C., J. A. Davis, R. E. Meyer, L. C. Messer, and T. J. Luben (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.

[136] Warren, J., M. Fuentes, A. Herring, and P. Langlois (2012). Bayesian spatial-temporal model for cardiac congenital anomalies and ambient air pollution risk assessment. *Environmetrics 23*(8), 673 – 684.

[137] Warren, J., M. Fuentes, A. Herring, and P. Langlois (2013). Air pollution metric analysis while determining susceptible periods of pregnancy for low birth weight. *ISRN Obstetrics and Gynecology 2013*, 1 – 9.

[138] Wikle, C. K. and L. M. Berliner (2005). Combining Information Across Spatial Scales. *Technometrics 47*(1), 80 – 91.

[139] Willocks, L., A. Bhaskar, C. Ramsay, D. Lee, D. Brewster, C. Fischbacher, J. Chalmers, G. Morris, and M. Scott (2012). Cardiovascular disease and air pollution in scotland: no association or insufficient data and study design? *BMC Public Health 12*, 227.

[140] Wiwanitkit, V. (2007). PM10 in the atmosphere and incidence of respiratory illness in Chiangmai during the smoggy pollution. *Stochastic Environmental Research and Risk Assessment 22*(3), 437 – 440.

[141] Wong, C.-M., N. Vichit-Vadakan, H. Kan, and Z. Qian (2008). Public Health and Air Pollution in Asia (PAPA): A Multicity Study of Short-Term Effects of Air Pollution on Mortality. *Environmental Health Perspectives 1*(1), 73 – 84.

[142] World Health Organisation (2014). Ambient (outdoor) air quality and health: Fact Sheet 313. http://www.who.int/mediacentre/factsheets/fs313/en/.

[143] Young, L. J., C. A. Gotway, J. Yang, G. Kearney, and C. DuClos (2009). Linking health and environmental data in geographical analysis: It's so much more than centroids. *Spatial and Spatio-temporal Epidemiology 116*(9), 1195 – 202.

[144] Yu, O., L. Sheppard, T. Lumley, O. J. Koenig, and G. G. Shapiro (2000). Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study. *Environmental Health Perspectives 108*(12), 1209–1214.