Estimating the effects of air pollution on human health in Greater Glasgow in space and time

Greg Jonathan Halbert

A Dissertation Submitted to the
University of Glasgow
for the degree of
Master of Science

School of Mathematics & Statistics

March 2013

Abstract

It is well documented that air pollution has an adverse effect on human health. With the increased risk of global warming, there has been an international effort to decrease emissions and pollution concentrations throughout the globe over the past sixty years, and these values are monitored by many laws and acts of governments. This thesis is a long term study of the effects of air pollution on the health of a Scottish population, specifically the incidence of respiratory disease cases in the Greater Glasgow and Clyde National Health Service (NHS) health board. As this is a long term study, the main points of interest are what effects pollution concentrations have on the hospitalisation counts of patients with respiratory disease on a yearly basis, and what other covariates, if any, have an effect on disease incidences. Furthermore, as this is a study in space and time we need to take into account any spatial and/or temporal correlation that may exist within the data. The study region is split up into 271 small areas based on population size and we evaluate what effect two specific pollutants, Nitrogen Dioxide (NO_2) and Particulate Matter (PM_{10}) have on respiratory disease across these areas. The rest of this thesis is structured as follows. Chapter One will present an introduction to the data and a literature review of the previous studies in this field. Chapter Two gives an outline of all of the statistical methods used throughout this study, including Poisson generalised linear models, diagnostic tests for overdispersion and spatial correlation, Bayesian models and conditional autoregressive models. Chapter three gives a description of all the data in the study and how it was obtained, as well as some pre-

liminary tables and plots. Chapter Four gives the results of all the purely spatial models discussed in Chapter Two. Chapter Five gives the results of the spatial-temporal health models where the entire space-time data set is modelled. Finally, Chapter Six presents an overall conclusion to the thesis, a discussion of any problems that occurred during this study, as well as what future work could be produced based on this study.

Acknowledgements

First of all I would like to thank my supervisor Dr Duncan Lee for all the hard work he did both with and for me. It is greatly appreciated and I hope I wasn't too big of a pain in the neck. I would also like to thank the Information Services Division (ISD) of NHS Scotland who provided the funding for my research to the University. Without their generous gift, this wonderful opportunity would not have been available to me.

I would like to thank my Mum and Dad for the wonderful support they gave me throughout my MSc, as well as the 23 and a half years prior to me starting my postgraduate degree, and also thanks to my brother for his support as well.

I also want to thank all of SYB who provided a great avenue to channel any stress and tension, as well as help clear my head when it was needed. Lastly, and by no means least, I would like to thank all the MSc crew of 420 (Stephen, Kathryn, Mhairi, Laura, Andisheh and Collette) who provided a lot of laughs, a great amount of help, but most importantly made sure the kettle was always full.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

It is well known and publicised that air pollution has adverse effects on the health of the population and the environment. The largest and most significant incidence of intense air pollution in the UK was the Great Smog in London of 1952. During the week of the 5th to the 9th of December, the smog got so thick that people could not see their own feet (Parliamentary Office of Science and Technology (2000)). The smog reached toxic levels as there had been no wind during that week. This caused the sulphur dioxide and smoke emissions from the city's factories, power plants, and domestic fireplaces burning cheap sulphurous coal to build up above the city instead of being blown away.

At the time, there had already been several smog/fog events (Davis et al. (2002)) in London. However, during the 1952 event hospital admissions greatly increased, as did the number of deaths. It is estimated that there were 4000 deaths above the normal mortality figures of London during the week of the smog, and it also took several months for London mortality figures to normalize after the event (Bell et al. (2004)).

After this event it was clear that air pollution can be very hazardous to health, and if left unattended can escalate to fatal levels very quickly. This

led to the government declaring the Clean Air act of 1956 (Clean Air Act (1956)), the main focus of this was on black smoke levels. This introduced "smoke control zones" in city centers where only smokeless fuels can be used. Another requirement of the act was that power plants had to be moved out of city centers and urban areas to more rural, less populated areas. Local authorities were allowed to designate how much smoke could be produced by an industrial site. Homes were also changed to be heated by cleaner coals, electricity and gas instead of the sulphurous cheap coal that was used previously. Due to this act, as well as the decline of the industrial sector in this country and the burning of cleaner more efficient fuels, there has been a decrease in smoke levels, and in parallel sulphur dioxide (SO_2) levels, of 90% compared to the smoke levels of the early fifties (Parliamentary Office of Science and Technology (2000)). The Clean Air Act has since been updated in 1968, and again in 1993.

Nowadays, smoke and SO_2 from factory emissions are not the main cause of the dangerous pollutants in the atmosphere, as they have been reduced. However, there has been a vast increase of road vehicles in the past 60 years and these are now the largest cause of air pollution in urban areas. Pollutants are split into two main groups, primary and secondary. Primary pollutants are direct emissions from a source such as cars, factories, fossil fuel power stations and homes. The most common harmful pollutants are different nitrogen oxides (such as NO_x and NO_2), volatile organic compounds (VOC), carbon monoxide (CO), carbon dioxide (CO_2), and fine particulates in the air. Particulates are small particles in the air, such as soot, dust and sea salt. One of the most highly monitored and regulated pollutants today are fine particles, or particulate matter (PM). The small particles in the air are formed from combustible sources, with one of the biggest sources of these being road traffic, reactions in the atmosphere of other gasses and pollutants forming secondary particles, and any small particles that may be floating in

the atmosphere. Examples of the latter include sand and dust, sea salt, plant matter such as seeds or pollen, or building materials from construction sites like sawdust and brick dust. Secondary pollutants are formed when primary pollutants mix in the atmosphere, such as ozone (O_3) at ground level, which occurs when nitrogen oxides, VOCs and sunlight react together.

Large particles in the air do not tend to enter the body as they are blocked in the nasal and throat passages by mucus and cilia. However, smaller particles that can get by these defences of the body are defined into four different categories;

- PM_{10} are particles that are less than 10 micrometers in aerodynamic diameter.
- $PM_{2.5}$ are particles that are less than 2.5 micrometers in aerodynamic diameter.
- PM_1 are particles that are less than 1 micrometers in aerodynamic diameter.
- Ultra fine particles (UFP) are particles that are less than 0.1 micrometers in aerodynamic diameter.

Depending on their size, they can have different effects on health (Laden et al. (2000)). PM_{10} can enter and settle in the bronchi and lungs, leading to breathing problems. $PM_{2.5}$ particles are more harmful as they are smaller particles made up of more hazardous toxic particles and small bits of metals in the air. As $PM_{2.5}$ contains smaller particles they can penetrate the lungs further than PM_{10}, into the gas exchange regions and enter the blood stream. Particles smaller than $PM_{2.5}$ can even enter into major organs. There are EU regulations that limit the level of PM concentrations in the atmosphere (Longhurst et al. (2009)). These objectives have been adopted and are enforced by the Air Quality Standards (Scotland) Regulations 2007. They state

that for the UK as a whole, a 24 hour mean PM_{10} concentration cannot exceed $50\mu gm^{-3}$ more than 35 times a year, and the yearly average of PM_{10} must be no higher than $40\mu gm^{-3}$, and these targets were to be met by the 31st of December 2004. There are also extra guidelines for Scotland, which state that, since 31st December 2010, the 24 hour mean PM_{10} concentration cannot exceed $50\mu gm^{-3}$ more than 7 times a year, and the yearly average must be no higher than $18\mu gm^{-3}$.

Nitrogen oxides are another common pollutant formed during combustible reactions that are very hazardous to health. Though nitrogen oxide (NO_x) is not harmful, when it is released into the atmosphere it oxidizes and becomes Nitrogen Dioxide (NO_2). NO_2 is toxic when inhaled, and can irritate the lungs as well as lower a body's resistance to infections such as flu. NO_2 is also closely linked to asthma in children (Gauderman et al. (2005)). The current regulations for NO_2 (enforced as from 31st December 2005) are that it should not exceed a mean value of $200\mu gm^{-3}$ in a 24 hour period more than 18 times in one year, and for the annual mean to be less than $40\mu gm^{-3}$.

To be more informative to the public, the level of each pollutant has been categorised into 10 bands of severity as approved by the committee on the medical effects of air pollution (COMEAP). Bands 1-3 are low air pollution, meaning that the effects of the pollutant is unlikely to be noticed by any individual, even one who is sensitive to air pollution or has respiratory problems. Bands 4-6 are moderate air pollution, which means most people should not feel any adverse effects, but sensitive individuals may notice mild effects. Bands 7-9 are classed as high, meaning that sensitive individuals may have strong reactions to the air, and should try to reduce their exposure. Asthma sufferers may also notice their inhalers do not have much effect. Band 10 is classed as very high.

The health outcome most often linked to air pollution is respiratory disease, and affects the lungs, bronchia, and surrounding tissue that are related to breathing. There are many different types of respiratory disease, ranging from the very light to the severe. One of the most common respiratory diseases that affects most people is the common cold. More life threatening examples are pneumonia and pulmonary embolism.

There have been many studies through the years investigating the short-term effect that air pollution has on human health. Numerous studies have been conducted in many countries and cities around the world. In the United States of America, there is the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) (Dominici et al. (2002)) that has investigated the short-term effects of air pollution on human health in 88 of the largest metropolitan ares in the US. In Europe, there is the Air Pollution and Heath - a European Approach (APHEA) study (Samoli et al. (2001)), which assesses the short term effect of air pollution on mortality and morbidity in 15 European cities. There are two main types of study of the effect of air pollution on health; short term studies and long term studies. Short term studies focus on the immediate effect that air pollution has, ie "what impact does high exposure to pollution over a couple of days have on health?". These kinds of studies look at the daily outcomes of health such as daily mortality or morbidity counts, and regress them against daily pollution levels, as well as other covariates of interest such as temperature.

Long term studies focus on the effect of prolonged exposure on health, i.e. "what effect does exposure to pollution over months and years have on health?". There are two types of long term study, cohort studies and small area ecological studies. In long term studies, counts of health outcomes from defined geographical areas over a pre-defined time period, are regressed against the air pollution concentrations for the same period of time, as well as

other covariates of interest such as socio-economical deprivation. Examples of these studies are Jerret et al. (2005), Maheswaran et al. (2005), Maheswaran et al. (2006), Elliot et al. (2007), Lee et al. (2009), Young et al. (2009), Haining et al. (2010) and Lee (2012).

Though there have been many studies on air pollution and health (for example (Elliot et al. (2007))), there have been few studies of air pollution and health data in Scotland, and even less of them have been long term studies. Only Lee et al. (2009) and Smith et al. (1987) are both long term studies of the effect of air pollution on health based in Scotland. Prescott et al. (1998) and Carder et al. (2008) are short term studies based on Scottish data, and Fairbarn & Reid (1958) is a short term study of the effect of air pollution on respiratory disease in the UK, including Scotland. Therefore in this thesis I intend to add to the limited body of evidence about the long term effects of air pollution on health in Scotland. The data used in this study are counts of the numbers of hospital admissions with a primary diagnosis of respiratory disease in the Greater Glasgow and Clyde NHS health board from 2002 to 2008. The health board is split up into 271 intermediate geographies (IG), which are small areal units designed for the distribution of small area statistics. They are based on population size (about 4000 people live in each one) and largely respect geographical boundaries (motorways, railways etc), and Scottish parliamentary constituency boundaries. More information can be found at http://www.scotland.gov.uk. The count data are the total number of hospital admissions with a primary diagnosis of respiratory disease for each of the 271 IGs within this study.

The remainder of this thesis is organised as follows. Chapter 2 will outline the statistical methods and theory used throughout this thesis. Chapter 3 will summarise the data graphically and numerically, while Chapter 4 will apply spatial regression methods to the data from each year separately. Chapter 5

will apply spatio-temporal models to the data, so that the overall effects of air pollution on respiratory health can be observed. Chapter 6 will conclude the thesis with a discussion of the methods used and any problems that arose during the study.

Chapter 2

Statistical background

2.1 Exploratory measure of disease risk

In this thesis the study region of Greater Glasgow is split into n non-overlapping small spatial units, and the number of disease cases observed in each small-area during a one year period is recorded. Therefore the disease data take the form of a count for each spatial unit, and should therefore be modelled by the Poisson distribution. See, for example, McColl (1995). This is because the Poisson distribution is primarily used to model the total number of events that occur in a fixed amount of time or space. Letting $\mathbf{Y} = (Y_1, \ldots, Y_n)$ denote the number of disease cases in each small-area, the likelihood function for Y_k is given by

$$f(Y_k; \mu_k) = \frac{e^{-\mu_k} \mu_k^{Y_k}}{Y_k!} \qquad \text{for } k = 1, \ldots, n. \tag{2.1}$$

The Poisson distribution makes the restrictive assumption that the mean and variance of Y_k are the same and equal to μ_k. The size and demographic structure of the population living in each small-area is different, and this should be accounted for when modelling $\mathbf{Y}$. This is achieved by calculating the expected number of disease cases in each small-area, which is denoted here by $\mathbf{E} = (E_1, \ldots, E_n)$. The expected number of cases in area k is calculated by splitting the population living in that area into strata based on their age and

sex, for example males 0-4, males 5-9, etc. Let n_{ki} be the number of people living in small-area k from stratum i, and r_i be the associated disease rate for that stratum from the entire study region. Then, the expected number of cases in area k is calculated as

$$E_k = \sum_{i=1}^{m} n_{ki} r_i. \tag{2.2}$$

A simple model for disease risk is given by

$$Y_k \sim \text{Poisson}(\mu_k = E_k R_k) \qquad \text{for } k = 1, \ldots, n \tag{2.3}$$

where the mean of Y_k is equal to the expected number of cases E_k multiplied by the disease risk R_k. Hence R_k denotes the overall risk of disease in area k, and its maximum likelihood estimate is given by

$$\hat{R}_k = \frac{Y_k}{E_k}. \tag{2.4}$$

This simple estimate of disease risk is also known as the Standardized Incidence Ratio (SIR), and a value of one corresponds to observing as many disease cases as you expect. Values greater than one denote unhealthy areas, for example, $\hat{R}_k = 1.1$ means that there were 10% more cases of respiratory admission in area k than were expected. Similarly, values less than one relate to healthy areas, with $\hat{R}_k = 0.8$ corresponding to 20% fewer admissions than expected from the population size and structure.

However, the estimate of R_k given by (2.4) is unstable, especially if the expected number of cases E_k is small. For example, if $E_k = 1$, then if you observe just two more cases than you expect (for example by chance), then you have a very extreme risk of 3. Therefore, an alternative model for Y_k is required, that does not produce such unstable estimates. This can be achieved by representing R_k as a linear combination of covariate risk factors, which has the advantage of using all the data points $\mathbf{Y}$ to estimate each

area's disease risk.

2.2 Simple regression models for R_k

A simple regression model for $\mathbf{Y}$ that represents the set of disease risks as a linear combination of covariates is a Poisson generalised linear model (GLM). The specific model used in this thesis is given by

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k), \\ \ln(R_k) &= \mu + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta}, \end{aligned} \tag{2.5}$$

where $\mathbf{x}_k^{\mathrm{T}} = (x_{k1}, \ldots, x_{kp})$ is the vector of covariate risk factors of interest and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the corresponding regression coefficients. The remaining coefficient μ is the intercept term, and the set of parameters are estimated using maximum likelihood. The effects of the covariates on the set of disease risks are measured by the parameter estimates $\hat{\boldsymbol{\beta}}$, and a 95% confidence interval for component β_i can be calculated as

$$\hat{\beta}_i \pm 1.96 \times \text{Standard Error}(\hat{\beta}_i) \tag{2.6}$$

However, the regression parameters are hard to interpret on this scale, as the data are being modelled by a log link function. Therefore, we transform the regression coefficients (and the confidence intervals) to the relative risk scale. The relative risk measures the percentage increase/decrease in the risk of disease given a specific increase in one of the covariates. For example, the relative risk associated with an increase in x_1 of ω units is given by

$$\begin{aligned} RR(\omega, \hat{\beta}_1) &= \frac{E_k \exp(\hat{\mu} + (x_{1k} + \omega)\hat{\beta}_1 + \sum_{j=2}^{p} x_{kj}\hat{\beta}_j)}{E_k \exp(\hat{\mu} + x_{1k}\hat{\beta}_1 + \sum_{j=2}^{p} x_{kj}\hat{\beta}_j)} \\ &= \exp(\omega\hat{\beta}_1) \end{aligned} \tag{2.7}$$

Confidence intervals on this transformed scale can be calculated by applying the exponential transformation $\exp(\omega.)$ to each end of (2.6). The choice of ω is somewhat arbitrary, but one approach is to use the standard deviation of each covariate, as it represents a realistic increase in its value. A relative risk of one means that the covariate has no effect on the disease data, while values greater than one suggest that increasing the covariate will increase the disease risk.

However, model (2.5) makes the following two limiting assumptions, which may not be realistic for the disease data analysed in this thesis.

1. The mean and the variance of each Y_k are equal.

2. Independence of $Y_1, \ldots, Y_n$.

The validity of these assumptions can be tested by examining the residuals, after model (2.5) has been fitted. The residuals we use are defined by

$$r_k = \frac{Y_k - E_k\hat{R}_k}{\sqrt{E_k\hat{R}_k}}$$

The first of these assumptions is that the mean and variance of Y_k will be equal, which is unlikely in spatial count data of this type. If the variance of Y_k is greater then this is known as overdispersion, while if the variance is smaller than the mean, it is known as underdispersion. To determine whether the mean and variance are equal we need to estimate the overdispersion parameter:

$$\phi = \frac{1}{n-p}\sum_{k=1} r_k^2. \tag{2.8}$$

where p is the number of parameters in the fitted model. If ϕ is equal to one, then the mean and variance are equal. However, if ϕ is greater than one, it shows evidence that the variance of the data is greater than the mean, and

there is overdispersion in the data. Similarly, the data will show evidence of underdispersion if ϕ is less than one.

As the data in this study relate to small spatial units, the assumption of independence may not be true. Instead, it is very likely that there will be high correlation between observations that relate to small-areas that are close to each other. To test if the data are independent, we calculate Moran's I statistic. Moran's I statistic calculates the strength (if any) of the correlation that exists in the data. Moran's I statistic is usually applied to the residuals after a regression model (such as (2.5)) has been fitted. For further information on Moran's I statistic, see Lawson (2009) or Moran (1950). In this case it is given by

$$I = \frac{n \sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{\sum_i (r_i - \bar{r})}. \tag{2.9}$$

Here, w_{ij} is a binary variable that defines whether areas (i, j) are neighbours. Two small-areas i and j are typically defined to be neighbours if they share a common border, in which case w_{ij} is equal to one. However, if they are not neighbours then w_{ij} will equal zero. If the value of Moran's I is close to one, then there is strong positive correlation in the residuals, i.e. the closer two areas are the more similar their values are. In contrast, if Moran's I is close to -1, then the data contain strong negative correlation. Finally, if Moran's I is close to zero, then there is no correlation and the data form a random spatial pattern, i.e. the data are independent. For example, if $I = 0.79$ then there is strong positive correlation in the data, but if $I = 0.09$ then there is very weak correlation. To test whether the value for Moran's I shows significant correlation, a permutation test can also be conducted. This involves calculating (2.9) for 10,000 sets of replicate independent data, which are permutations of the original data set (i.e. you randomly allocate each observation to an area). If the calculated p-value of this permutation test is less than 0.05 there is significant evidence of spatial correlation in the data,

which means that the assumption of independence is not valid. If evidence of spatial correlation exists, then the model defined by (2.5) is not a good fit for the data. To analyse the data, taking into account any spatial correlation between small-areas, Bayesian methods are typically used. An advantage of Bayesian method is that we can fit the spatial data using a prior distribution which will explain how we believe the spatial correlation will behave in the model and account for this to give more accurate results. For an example of a comparison of frequentist and Bayesian spatial methods, see (Ismaila et al. (2007)).

2.3 Introduction to Bayesian methods

While there are frequentist methods to model spatial data, Bayesian methods are more commonly used to model spatial data, so we provide a brief introduction here. In general, let us define a data vector as $\mathbf{Y} = (Y_1, \ldots, Y_n)$, which depends on parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The likelihood function describes the information in $\mathbf{Y}$ about $\boldsymbol{\theta}$, which can be expressed as

$$f(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i|\boldsymbol{\theta}), \tag{2.10}$$

provided $Y_1, \ldots, Y_n$ are assumed to be independent. In a Bayesian analysis you additionally specify a prior distribution $f(\boldsymbol{\theta})$, to define how we believe that the parameters will behave before the data have been observed. Once $f(\mathbf{Y}|\boldsymbol{\theta})$ and $f(\boldsymbol{\theta})$ have been defined, they can then be used to calculate the posterior distribution, which describes the behavior of the parameters after the data have been observed. The posterior distribution is the combination of the prior information and the likelihood function, and using Bayes theorem is given by

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{Y}) &= \frac{f(\boldsymbol{\theta})f(\mathbf{Y}|\boldsymbol{\theta})}{f(\mathbf{Y})}, \\ &\propto f(\boldsymbol{\theta})f(\mathbf{Y}|\boldsymbol{\theta}). \end{aligned} \tag{2.11}$$

The simplification on the second line to remove $f(\mathbf{Y})$ can be made as it is not dependant on $\boldsymbol{\theta}$. If the posterior distribution is a standard distribution, then inference about $\boldsymbol{\theta}$ (e.g. mean, 95% credible intervals) is straightforward to obtain. This situation occurs when the posterior distribution is from the same distributional family as the prior distribution. When this happens $f(\boldsymbol{\theta})$ is called a conjugate prior. However, if $f(\boldsymbol{\theta}|\mathbf{Y})$ is not a standard distribution then we can make inference about it by simulating random draws from the posterior distribution.The most common way of generating these random numbers is using Markov Chain Monte Carlo (MCMC) methods, and a brief description of the two most common methods are given below.

The first is the Gibbs sampler (Geman & Geman (1984)), which generates a sequence of samples from the conditional distribution of θ_i given all other values for $\boldsymbol{\theta}$ and the data $\mathbf{Y}$. We set an initial state to be $\boldsymbol{\theta}^{(\mathbf{0})} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$, which is randomly generated from the sample space. We then repeat the following steps for a large number of iterations, say 10,000. At iteration t, we sample $\theta_1^{(t)}$ from its conditional distribution $f(\theta_1^{(t)}|\theta_2^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y})$, which is a proper distribution. That is, we generate θ_1 from its full conditional distribution given the current values of the remaining parameters and the data. This step is repeated for each θ_i in turn, before moving on to iteration $t + 1$.

The other most commonly used MCMC method is the Metropolis-Hastings algorithm based on (Hastings (1970)), which is used when the full conditional distribution of a parameter is not proper. We set an initial state of $\boldsymbol{\theta}^{(\mathbf{0})} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ which is randomly generated from the sample space as before. When $f(\theta_1^{(t)}|\theta_2^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y})$ is not proper, we generate a proposed value θ_1' from a proposal distribution $q(\theta_1', \theta_1^{(t)})$. This new sample is accepted as the next value of the chain $\theta_1^{(t+1)} = \theta_1'$ if we sample α from a uniform distribution U(0,1) and it meets the following criteria

$$\alpha < \frac{f(\theta_1'|\theta_2^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y}) q(\theta^{(t)}, \theta')}{f(\theta_1^{(t)}|\theta_2^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y}) q(\theta', \theta^{(t)})}. \tag{2.12}$$

If α does not satisfy this requirement, then the sample value is the same as the previous state, $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$.

Given you've generated 10,000 random draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(10,000)}$ from $f(\boldsymbol{\theta}|Y)$ using either of the two previous methods, then posterior inference becomes straightforward. The posterior mean and variance can be calculated as

$$\begin{aligned} \mathbb{E}[\theta_i|\mathbf{Y}] &= \frac{1}{10,000} \sum_{t=1}^{10,000} \theta_i^{(t)}, \\ \mathrm{Var}[\theta_i|\mathbf{Y}] &= \frac{1}{(10,000-1)} \sum_{t=1}^{10,000} (\theta_i^{(t)} - \bar{\theta})^2, \text{ where } \bar{\theta} \text{ is } \mathbb{E}[\theta_i|\mathbf{Y}]. \end{aligned}$$

In this thesis the data vector $\mathbf{Y}$ are the counts of the number of disease cases in each small area k as before. The likelihood function is

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k), \\ \ln(R_k) &= \mu + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta} + \phi_k, \end{aligned} \tag{2.13}$$

where the parameters are $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}, \boldsymbol{\phi})$, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are the random effects to model the spatial correlation in the data. We define a prior distribution on the parameters as

$$f(\boldsymbol{\theta}) = f(\mu, \boldsymbol{\beta}, \boldsymbol{\phi}) = f(\mu) f(\boldsymbol{\beta}) f(\boldsymbol{\phi}), \tag{2.14}$$

where we assume independent priors. Details of the priors are given in the next two sections.

2.4 Introduction to conditional autoregressive models

The random effects $\boldsymbol{\phi}$ model any spatial correlation and overdispersion in the data. The most common model for $\boldsymbol{\phi}$ is a conditional autoregressive (CAR) model, based on the information in Banerjee et al. (2004). We wish to specify a joint prior distribution of $f(\boldsymbol{\phi})$ for the random effects that induces spatial correlation. However, CAR models instead specify a conditional distribution on each individual ϕ_k $f(\phi_k|\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$. This simplifies to $f(\phi_k|$a set of neighbours) based on which other small-areas ϕ_k is neighbours with. In this thesis neighbours are two areas that share a common border. This neighbourhood information is contained in W, a binary $n \times n$ neighbourhood matrix where

$$w_{jk} = \begin{cases} 1 & \text{if area } j \text{ shares a common border with area } k, \text{ denoted as j} \sim \text{k} \\ 0 & \text{otherwise} \end{cases}$$

The intrinsic CAR model was proposed by Besag et al. (1991). The full conditional distributions of the spatial effect ϕ_k is given by

$$f(\phi_k|\phi_{-k}) \sim N\left[\frac{\sum_{j=1}^{n} w_{jk}\phi_j}{\sum_{j=1}^{n} w_{jk}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{jk}}\right], \quad \text{for } k = 1, \ldots, n \quad (2.15)$$

$$f(\phi_k|\phi_{j\sim k}) \sim N\left[\frac{1}{n_k}\sum_{j\sim k}\phi_j, \frac{\tau^2}{n_k}\right], \quad (2.16)$$

where n_k is the number of neighbours each small-area has.

However, problems exist within this type of CAR model. The intrinsic CAR model is only appropriate if very strong spatial correlation exists within the data as the single parameter τ only models the variation amongst the random effects and does not control the strength of the spatial correlation.

Another problem is that the distribution $f(\phi_1, \ldots, \phi_n)$ corresponding to the full conditional distributions of ϕ_k is improper. This is because the precision matrix $\mathbf{P} = \frac{1}{\tau^2}(D - W)$, where D is a $n \times n$ diagonal matrix with the number of neighbours each small-area k has, is not invertible. To make $\mathbf{P}$ invertible and make the CAR model proper, a parameter ρ can be added to the model which controls the strength of the spatial correlation and will make the distribution of $f(\phi)$ proper.

The conditional distribution for ϕ_k in the proper CAR model (Cressie (1993)) with the added parameter ρ is given by

$$f(\phi_k|\phi_{-k}) \sim N\left[\rho\frac{\sum_{j=1}^{n} w_{jk}\phi_j}{\sum_{j=1}^{n} w_{jk}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{jk}}\right], \quad \text{for } k = 1, \ldots, n \quad (2.17)$$

$$\sim N\left[\rho\frac{1}{n_k}\sum_{j\sim k}\phi_j, \frac{\tau^2}{n_k}\right] \quad (2.18)$$

In (2.18), the precision matrix $\mathbf{P} = \frac{1}{\tau^2}(D - \rho W)$ is invertible if $\rho \in [0, 1)$. If $\rho = 1$ then our model reverts back to the intrinsic model above. If ρ is close to 1, then there is strong spatial correlation. However, if there is weak correlation ρ will be close to zero and if $\rho = 0$, then the random effects are independent.

2.5 Introduction to autoregressive (AR) processes

We have looked at methods that will take into account any spatial correlation that may exist within the data, but we have not looked at any similarities that may exist in the data through time. The data we have are the number of hospital admissions of respiratory disease for each small area for the seven years of interest in our study. Therefore it is safe to assume that some of each small area's admission counts can be explained by the previous years values. One time series model that takes into account previous values is the autoregressive process (AR) of order j. As our study length is only seven years, we will use an AR process of order 1, which in generic notation can be written as

$$X_t = a_1 X_{t-1} + Z_t. \tag{2.19}$$

Here, X_t is the number of admissions at time t, X_{t-1} is the number of admissions at the previous time, a_1 is the lag one coefficent and Z_t is white noise. We will use this process to model the temporal correlation in the data. For further information see Pandit & Wu (1983).

2.6 Spatial-temporal models

The final part of this thesis models the relationship between air pollution and health in space and time, so we will need a model that accounts for overdispersion and spatial correlation within the data, as well as any underlying temporal correlation within the data. To do this we are going to use a modified version of the model proposed by Knorr-Held (2000). In his paper, he proposed a main effects model with no covariates, and an interaction term. Letting R_{kt} be the risk in area k and time period t, Knorr-Held (2000) models this by

$$R_{kt} = \mu + \alpha_t + \gamma_t + \theta_k + \phi_k + \delta_{kt}. \tag{2.20}$$

In the above model, μ is the intercept term and

- $f(\alpha_t) \sim N(\alpha_{t-1}, \sigma_\alpha^2)$ is an AR process to model the temporal correlation in the data.
- $f(\gamma_t) \sim N(0, \sigma_\gamma^2)$ models independent errors over time.
- $f(\phi_k) \sim N\left(\frac{1}{n_k}\sum_{j\sim k}\phi_j, \frac{\tau^2}{n_k}\right)$ models the strong spatial correlation that exists in the data. This is the same as the intrinsic CAR given in section 2.4.
- $f(\theta_k) \sim N(0, \sigma_\phi^2)$ models independent errors in space.
- δ_{kt} models any interaction effect between space and time.

Knorr-Held then proposes the following four types of interaction. Type I interaction is if there are independent interactions in space and time. Type II interaction is when we have a correlated interaction in time, but not in space. Type III interaction is similar, as it has a correlated interaction in space but not in time. Type IV interaction is when there is a correlated interaction in both space and time.

Based on the format of our data, it is fair to assume that small areas near each other may have similar disease risk, and we have already proposed ways to model this in a purely spatial context. It is also fair to assume that disease risk will be similar in consecutive years of the study. We therefore use a simplified version of Knorr-Held's model, given by

$$\begin{aligned} Y_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\ \ln(R_{kt}) &= \mu + \mathbf{x}_{kt}^{\mathrm{T}}\boldsymbol{\beta} + \alpha_t + \phi_k. \\ \alpha_t &\sim N(\alpha_{t-1}, \delta_\alpha^2) \end{aligned}$$

In the above model, μ is the intercept term, $\mathbf{x}_{kt}^{\mathrm{T}}$ are the covariates of interest, $\boldsymbol{\beta}$ are the coefficient terms, α_t models the temporal effects and ϕ_k is the proper CAR model that models the overdispersion and spatial correlation in the data given in equation (2.18).

Chapter 3

Data and Descriptive Statistics

This chapter describes the data used for this thesis and presents spatial plots and descriptive statistics tables.

3.1 Health

The health data analysed in this thesis are all hospital admissions of both male and female patients of all ages diagnosed with a respiratory disease in the greater Glasgow and Clyde Health board between 2002 and 2008. This data was obtained from the Information Services Division (ISD) Scotland. This health board is split up into 271 intermediate geographies (IG). Each count is of admission and discharge to hospital, so one patient who is admitted to hospital then transferred to another consultant or hospital before being discharged is only counted once. However, if a patient is admitted to hospital, discharged and then re-admitted to hospital in the same year then that is counted as two separate admissions. Respiratory disease is defined using the International Classifications of Disease Volume 10 (ICD10), under codes J00 to J99 and R09.1. The data are obtained from the Scottish Neighbourhood Statistics (SNS) database, which is available online at http://www.sns.gov.uk/.

We will first look at the number of admissions in each year. Table 3.1 shows the lowest, median and highest numbers of observed admissions in each year of the study.

Table 3.1: Lowest, median and highest number of hospital admissions in each year.

Year	Lowest	Median	Highest
2002	14	58	162
2003	13	58	176
2004	8	57	171
2005	13	65	181
2006	13	63	181
2007	15	70	194
2008	10	75	208

From Table 3.1 we can see that the highest number of admissions seems to increase throughout time. In 2002 the highest number of admissions to hospital was 162, then 3 years later this figure has risen to 181, and in 2008 it has went up again to 208 patients. There is also evidence of an increase in the medians as well. In 2002 the median is 58, and it has increased to 65 in 2005, and increases again to 75 in 2008. The lowest numbers of hospital admissions seems to increase and decrease with no real pattern through the study period.

To compare the risk of respiratory disease admission over the health board we looked at the standardized incidence ratio (SIR) in each year of the study. The SIR for each IG was calculated by dividing the number of hospital admissions observed by the expect number of hospital admissions for that IG (see equation (2.2)). The expected number of hospital admissions were calculated using rates of hospital admission for the whole of Scotland. We use the SIR instead of the number of admissions as this will correct for differences in pop-

ulation size and demographic structure. This is because we would expected an area with a larger population to have a larger number of admissions so to compare the observed number of cases to a less populated area would be unfair. We will again look at a table of the lowest, median and highest values within each year, as well as spatial plots of the SIR for the beginning, middle and end of the study period, ie 2002, 2005, and 2008.

Table 3.2: Lowest, median and highest SIR in each year.

Year	**Lowest**	**Median**	**Highest**
2002	0.2765	0.8513	2.042
2003	0.2097	0.8425	1.833
2004	0.1914	0.8001	1.897
2005	0.3251	0.8621	2.021
2006	0.2781	0.8026	1.640
2007	0.3009	0.8891	1.809
2008	0.2328	0.9331	2.269

In Table 3.2 the median risk level in each year looks to be quite consistently around 0.85, except for an increase in 2008 to a median value of 0.9331. The highest risk level is also in 2008, with a value of 2.269. There does not seem to be any other pattern in the maximum SIR values as they seem to increase and decrease from approximately 1.8 to 2.0 quite regularly, with the exception of 2006 which has the SIR rate of 1.64. From Tables 3.1 and 3.2 there is evidence of 2008 being the year of most admission to hospital with respiratory disease, as well as the highest risk of hospital admission, and there seems to be an overall increase in these figures through the study period.

Figures 3.1, 3.2 and 3.3 all seem to show that the areas with the highest risk of respiratory disease are the small areas all clustered together, which correspond to the deprived east end of Glasgow. The areas of lowest SIR are the large areas outside of the city, and the west end of Glasgow. These

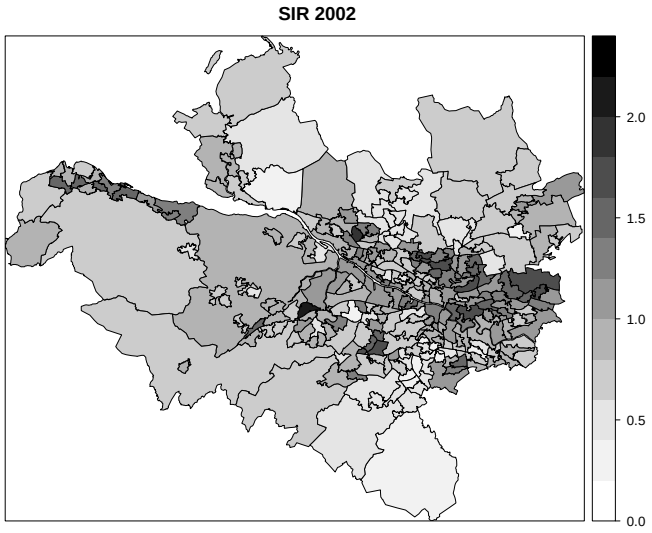

Figure 3.1: Spatial plot of SIR in 2002 in Greater Glasgow and Clyde Health Board.

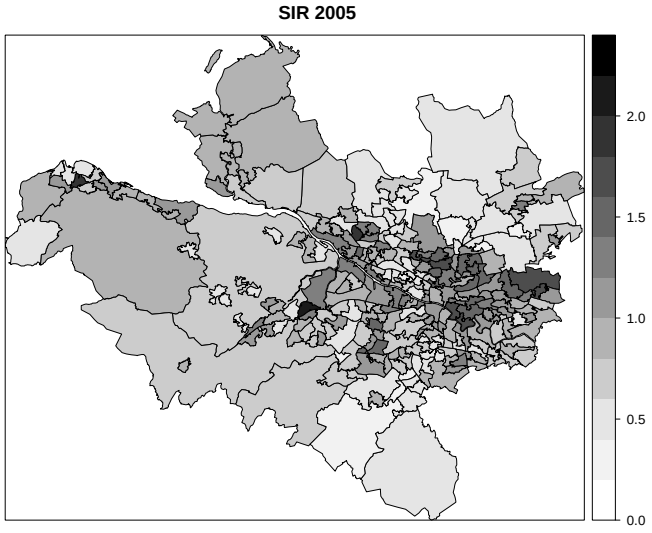

Figure 3.2: Spatial plot of SIR in 2005 in Greater Glasgow and Clyde Health Board.

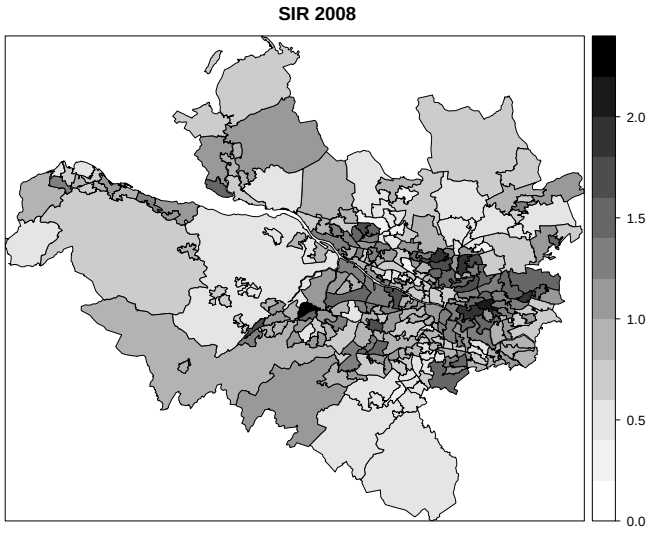

Figure 3.3: Spatial plot of SIR in 2008 in Greater Glasgow and Clyde Health Boadr.

seem pretty consistent throughout the study period. We can therefore conclude that there is evidence that the most at risk areas of respiratory disease in the study are the residential areas within Glasgow and this is consistent throughout the study period.

3.2 Air Pollution

We have obtained air pollution data for a number of pollutants recorded in the greater Glasgow and Clyde area. We analyse the pollutants from the previous year to the hospital admissions, as exposure to pollution is unlikely to have an immediate effect. It is more likely that air pollution will cause damage over time. The air quality data were obtained from the Department

for Environment Food and Rural Affairs (DEFRA) website. Dispersion models were used to estimate pollution levels at one kilometer intervals, and the resulting values were scaled using the much smaller number of air quality monitoring sites. To calculate a pollution value for each IG within the study region, we take the median value of the 1km modelled estimates within each IG. For small areas that do not contain any of these estimates, we use the value that is closest to it. There are only three pollutants throughout the $2001-2007$ study period of interest. Of these three we are going to focus on Nitrogen Dioxide (NO_2) and particles less than 10μm in diameter (PM_{10}). The third pollutant is Nitrogen monoxide, which has a very strong correlation with Nitrogen Dioxide so it is not included in the study.
To investigate the change in pollution values through time, Tables 3.3 and 3.4 show the lowest, median and highest pollution concentration values for both Nitrogen dioxide and PM_{10}.

Table 3.3: Lowest, median and highest Nitrogen Dioxide concentration in each year.

Year	**Lowest**	**Median**	**Highest**
2001	6.200	28.30	43.40
2002	6.033	27.54	42.23
2003	5.878	26.83	41.14
2004	3.080	18.90	38.30
2005	2.970	18.60	37.70
2006	3.402	14.70	34.76
2007	3.277	14.04	33.29

In Table 3.3 there is evidence of a decrease in Nitrogen Dioxide concentration throughout the study period. The largest values of NO_2 concentration is 43.4 in 2001. This value decreases slightly each year to 33.29 in 2007. There is also a similar pattern in the median values, with a high of 28.3 in 2001 decreasing to 26.83 in 2003 then a large drop to 18.9 in 2004. In 2006

Table 3.4: Lowest, median and highest PM_{10} concentration in each year.

Year	Lowest	Median	Highest
2001	13.10	17.30	20.5
2002	12.80	16.90	20.03
2003	12.50	16.50	19.56
2004	10.20	15.00	21.9
2005	10.10	15.00	21.7
2006	9.980	13.88	20.26
2007	9.908	13.71	19.89

the median drops quite low to 14.7, then down to the lowest median value of 14.04 in 2007. The lowest value of NO_2 concentration is 2.97 in 2005, and the largest low value is 6.2 in 2001. Table 3.4 also shows evidence that a slight decrease in air pollution, in this case for PM_{10}, may exist. The minimum value of PM_{10} concentration in 2001 is 13.1, which decreases slightly each year to 9.908 in 2007. The median values also decrease gradually every year from 17.3 in 2001 to 13.71 in 2007, with the exception of 2005 and 2006 which both have a concentration value of 15. The maximum values of PM_{10} decrease very slightly from 20.5 to 19.56 between 2001 and 2003, before a slight increase to 21.9 in 2004, then decreasing down to a value of 19.89 in 2007. We can therefore conclude that there seems to be evidence of a decrease in air pollutants from 2001 to 2007. To look for pollution changes within IGs through time Figures 3.4 to 3.9 present the spatial pattern of each pollutant for 2001, 2004 and 2007.

The spatial plots for PM_{10} in 2001 and 2004 show that the highest areas of pollution are all centered around Glasgow City Centre, and dilute as you travel further from Glasgow itself. In 2007 the PM_{10} seem very low throughout the entire health board, especially in the city of Glasgow. The plots of NO_2 show a very similar decrease through time as PM_{10}. However, in both

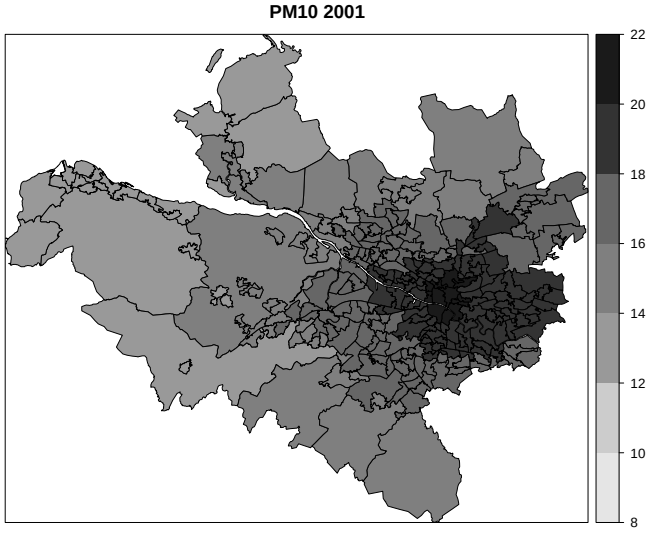

Figure 3.4: Spatial plot of PM_{10} in 2001 in Greater Glasgow and Clyde Health Board.

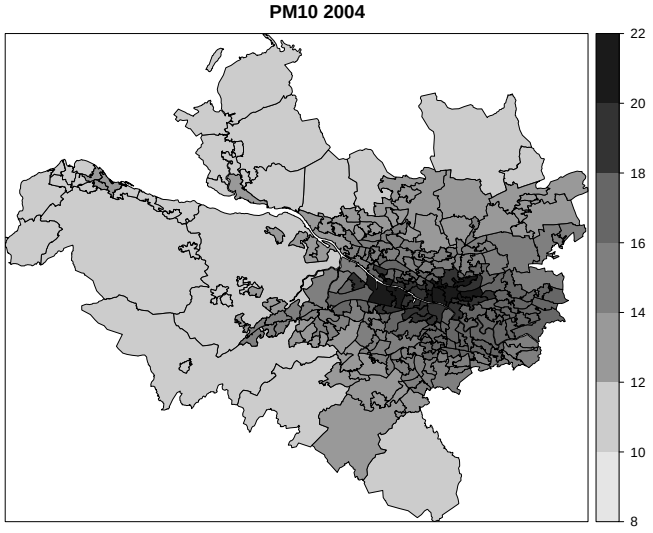

Figure 3.5: Spatial plot of PM_{10} in 2004 in Greater Glasgow and Clyde Health Board.

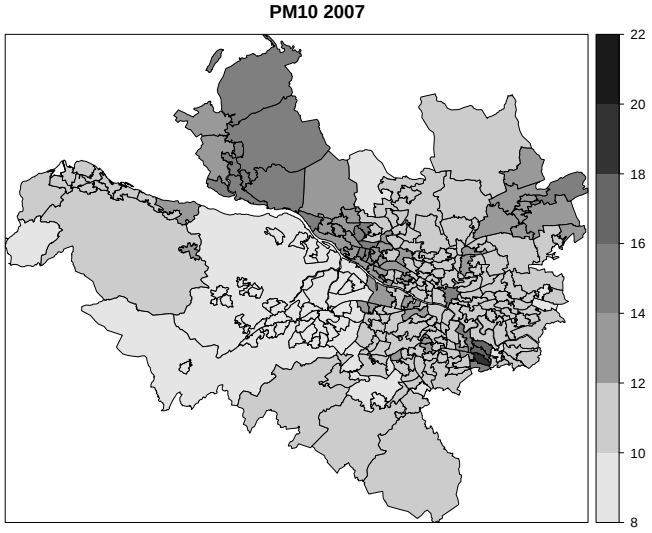

Figure 3.6: Spatial plot of PM_{10} in 2007 in Greater Glasgow and Clyde Health Board.

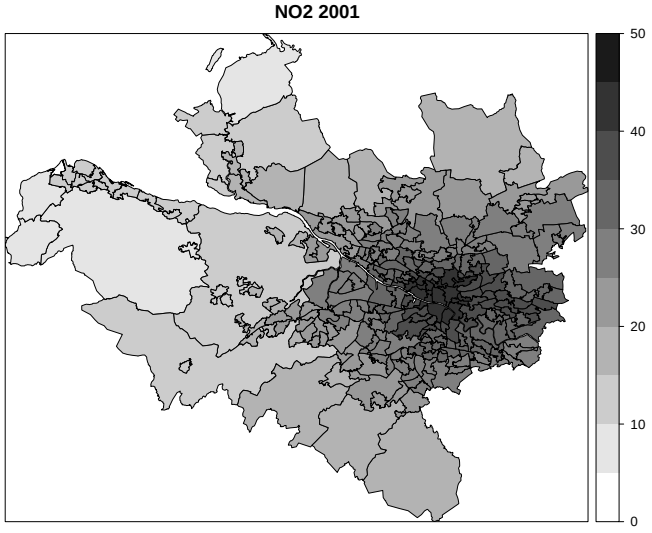

Figure 3.7: Spatial plot of NO_2 in 2001 in Greater Glasgow and Clyde Health Boadr.

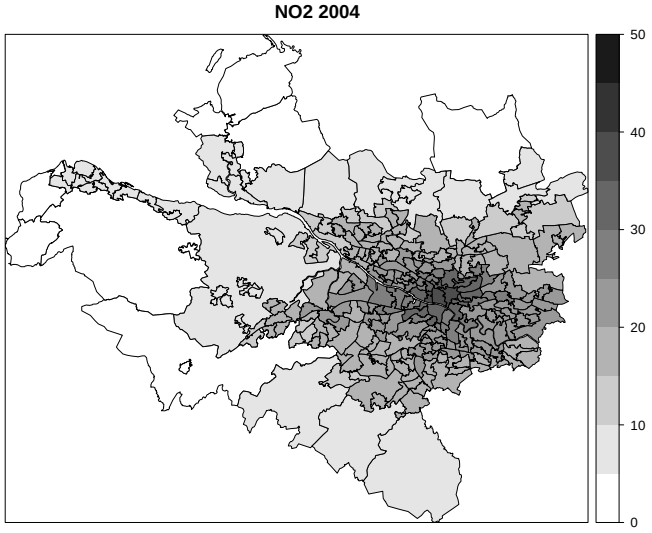

Figure 3.8: Spatial plot of NO_2 in 2004 in Greater Glasgow and Clyde Health Board.

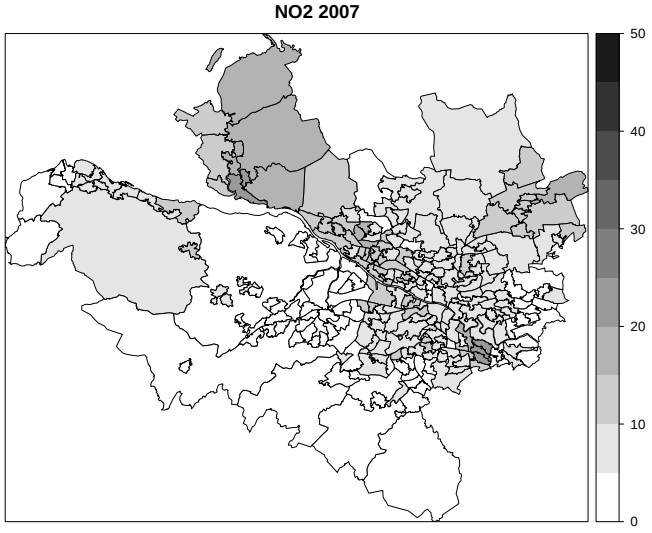

Figure 3.9: Spatial plot of NO_2 in 2007 in Greater Glasgow and Clyde Health Board.

Figures 3.6 and 3.9 seem to show a slight increase in pollution levels by 2007 in the north west of the study region.

3.3 Covariates

The covariates discussed in this section are any other covariates that we feel could have an important effect on hospital admission risk for respiratory disease. All of these covariates were obtained from the SNS website as before.

One of the covariates of interest we are looking at is the percentage of the population who smoke within each IG. These figures were calculated from the 2001 UK census and the 2003/04 Scottish Household Survey. As such these smoking data are only really appropriate for these years and do not take into account effects on the years from 2005 onwards. For example, there may be a lower percentage of smokers and a different spatial pattern after 2006 when the public smoking ban came into effect. The lowest percentage of the population who smoke within each IG is 10.5%, the highest is 52.4% and the median is 29.6%.

The plot of smoking against the SIR of 2003 shown in Figure 3.10 shows evidence of a positive linear relationship. The correlation coefficient for smoking against risk for 2003 is 0.759. This also shows strong evidence of a relationship between smoking and hospitalisation. The plot for 2004 also shows evidence of this relationship. Therefore there is evidence that there is a higher risk of hospitalisation with respiratory disease if more of the population smoke within each small area. Figure 3.12 shows that the highest concentration of the percentage of the population that smoke is in the small areas within the east end of Glasgow.

The next covariate of interest is the ethnic background of each IG. This is

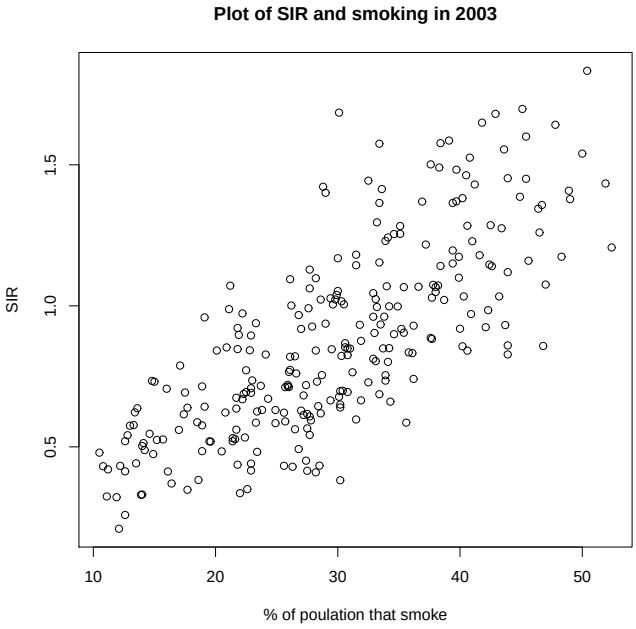

Figure 3.10: SIR and smoking 2003

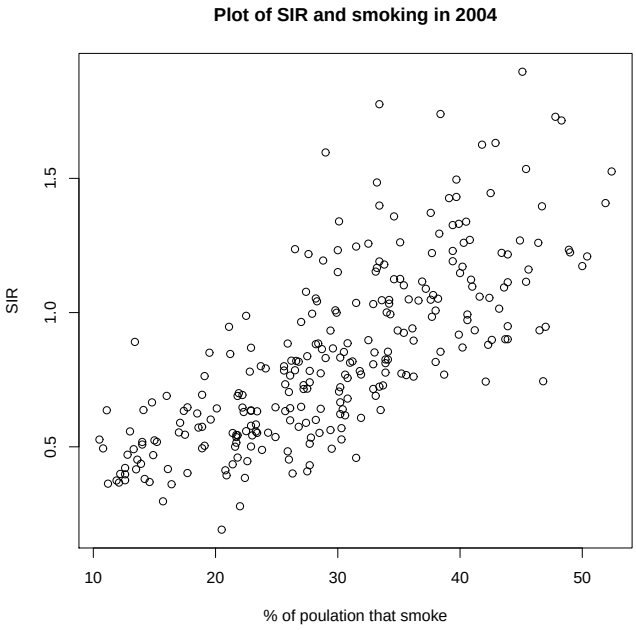

Figure 3.11: SIR and smoking 2004

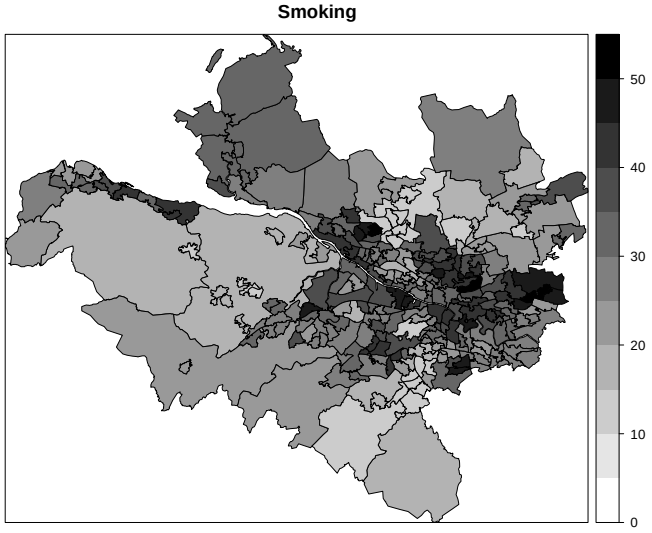

Figure 3.12: Spatial plot of smoking population in Greater Glasgow and Clyde Health Board %

measured by the percentage of children from ethnic minorities in each Intermediate Geography, e.g white or non-white. We have this data for the years 2004-2008. Table 3.5 shows the lowest, median, and highest percentage of ethnic children in each IG every year, however the summary statistics showed that the lowest values and median values are very close to each other, suggesting that most of these 271 areas have very small ethnic populations. To evaluate this further we have added the 75% quantile as well to get a better idea of how skewed the data are. Table 3.5 below shows that the largest 75% quantile value of ethnic children is 11.88% in 2008. This shows that the data are very skewed, ie the majority of the population is white. There is evidence of increases in the ethnic population between 2004 and 2008 in the median and 75% quartile values of Table 3.5, with the median gradually increasing from 3.1% in 2004 to 3.92% in 2007, before a slightly jump to 4.47% in 2008, and the 75% quartile increasing from 8.18% to 11.88%. There is a drop in the highest ethnic populated area from 83.7% in 2004 to 80.52% in 2005,

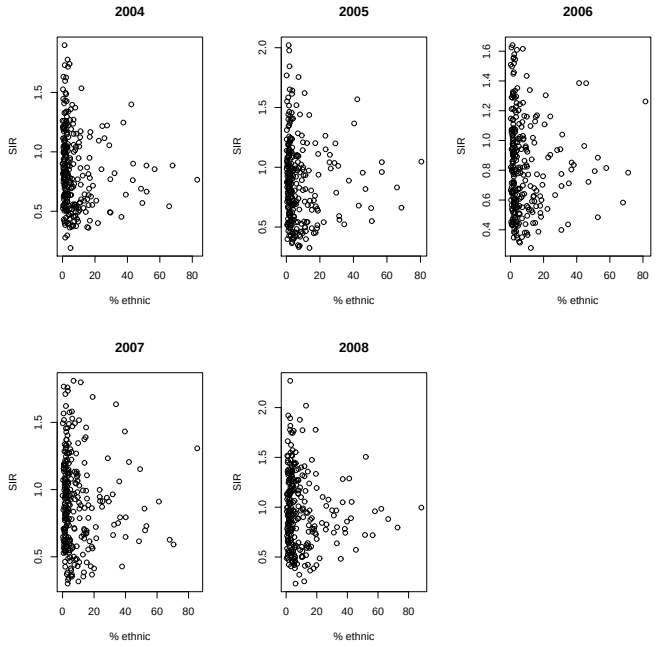

Figure 3.13: Plots of SIR and % of ethnic children

which then increases up to 88.31% in 2008. There does not seem to be an obvious pattern in the lowest percentile. Figure 3.13 shows that the ethnic population percentages are skewed to the right.

Table 3.5: Lowest, median, 75% quartile and highest % ethnic children in each year.

Year	Lowest	Median	75% Quartile	Highest
2004	0.00	3.10	8.180	83.70
2005	0.17	3.30	9.120	80.52
2006	0.14	3.49	9.805	81.49
2007	0.00	3.92	10.61	85.53
2008	0.32	4.47	11.88	88.31

We also have obtained a measure of deprivation, in this case the median price of a house in each IG. Median house price was chosen as it is not as highly correlated with smoking as the other available measures of deprivation

are. The mean correlation of smoking and house price is -0.7049481. Median house price is used as a proxy measure of overall deprivation, because poorer people typically smoke more, drink more, do not exercise etc, compared to rich people. We have this information for every year of the study period. We chose Median house price over using the Scottish Index of Multiple Deprivation (SIMD) as we felt median house price is a more informative variable and is in keeping with the rest of the data in this study. Figure 3.14 shows the median house price plotted against the SIR for 2002.

Table 3.6: Lowest, median and highest median house price (£'s) in each year.

Year	**Lowest**	**Median**	**Highest**
2002	28000	60000	262500
2003	26500	72000	298500
2004	34000	85180	317500
2005	41000	95000	318000
2006	41250	105600	352900
2007	57200	122000	430000
2008	50000	122000	372800

Table 3.6 shows that there seems to be an overall general increase in house price each year for the median and highest values. This is probably due to inflation. Figure 3.14 suggests that there could be a negative linear relationship between house price and SIR exists, ie as price increases, respiratory disease admissions decrease. There is evidence of this relationship for every year of the study.

Figure 3.14: Plot of SIR and median house price (£'s) 2002

Chapter 4

Spatial health models

We will will first investigate the effect of the covariates on hospital admission with respiratory disease using Poisson generalized linear models which ignore the possibility of spatial correlation and overdispersion. The models are fitted separately for each year of the study.

4.1 Poisson Generalised Linear Models

We first fitted the following Poisson glm separately for each year of the study using maximum likelihood estimation;

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k), \\ \ln(R_k) &= \mu + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta}, \end{aligned} \tag{4.1}$$

where the covariates $\mathbf{x}_k^{\mathrm{T}}$ are smoking prevalence, the log transformation of the median house price which has been applied to make the data easier to interpret, the percentage of children from ethnic minorities, and either PM_{10} or NO_2. Unfortunately we do not have data on smoking prevalence for every individual year, as it only relates to the years 2003 and 2004. However, we use this covariate for every year of the study to asses the effect of smoking on respiratory disease and to ensure the effects of air pollution are compa-

rable across all seven years of the study. Similarly, we only have data on the proportion of ethnic children from 2004 onwards, so for 2002 and 2003 we will use the 2004 values. As the ethnic population is very small and the corresponding percentages are very close to zero, we have applied a log transformation to this variable. To eliminate the problem of calculating the logarithm of 0, we have added a small constant of 0.05 to every value. The following tables show the relative risks for each pollutant in every year, which were calculated using equations (2.6) and (2.7). The ω increase used in these calculations is the standard deviation of the related covariate.

Table 4.1: Table of relative risks, 95% Confidence intervals and ω for PM_{10}.

Year	**Relative risk**	**95% confidence interval**	ω
2002	1.024	(1.003, 1.045)	1.986
2003	1.042	(1.020, 1.064)	1.986
2004	1.086	(1.062, 1.111)	1.986
2005	1.044	(1.028, 1.060)	1.986
2006	1.057	(1.041, 1.074)	1.986
2007	1.060	(1.040, 1.080)	1.986
2008	1.047	(1.028, 1.067)	1.986

Table 4.2: Table of relative risks, 95% Confidence intervals and ω for NO_2.

Year	**Relative risk**	**95% confidence interval**	ω
2002	1.017	(1.001, 1.033)	7.032
2003	1.031	(1.014, 1.048)	7.032
2004	1.071	(1.053, 1.089)	7.032
2005	1.060	(1.040, 1.080)	7.032
2006	1.075	(1.053, 1.096)	7.032
2007	1.076	(1.049, 1.102)	7.032
2008	1.089	(1.062, 1.117)	7.032

Table 4.1 shows that PM_{10} has a significant effect on the risk of respiratory disease for every year of the study, as none of the confidence intervals contain the null risk of one. There is therefore substantial evidence that PM_{10} is bad for health. The value of 1.986 for ω was calculated by taking the mean of the standard deviations of PM_{10} from 2002 to 2008. The relative risk of PM_{10} is lowest in 2002 and 2003, as for an ω increase in PM_{10} concentration there is a 2.4% and 4.2% increase in risk of respiratory disease respectively. The year of the highest relative risk of respiratory disease for an increase in PM_{10} of ω units is in 2004, with an increase of 8.6% and a confidence interval between 6.2% and 11.1%. Overall, there seem to be no obvious pattern in relative risk through time, but it does have a significant effect every year of the study, and has an overall average increase of 5.1%.

Table 4.2 shows that NO_2 has a significant effect on the risk of respiratory disease for every year of the study. The ω value was again calculated using the mean of the standard deviation of the covariate. In 2002 there was a 1.7% increase in relative risk per ω increase in NO_2, with a confidence interval of between a 0.1% and 3.3% increase. In 2003 the risk of NO_2 has increased to 3.1%, and the 95% credible interval has increased to between a 1.4% and a 4.8% increase. Both 2002 and 2003 have the smallest effect on respiratory disease as before for PM_{10}. Table 4.2 shows that 2004 has a mean increase of 7.1%, and the increase could be anything between a 5.3% increase and a 8.9% increase in relative risk. The years 2005 to 2008 show an increase in risk from year to year. In 2005 the relative risk of respiratory disease by NO_2 is 6%, which is smaller than 2004. The relative risk of 2006 then increases up to 7.5%, then increases to 7.6% in 2007, and then increases again in 2008 to 8.9%. The pattern for NO_2 and PM_{10} are similar for the first three years of the study, ie from 2002 to 2004. The relative risks for 2002 and 2003 are both very small, then there is a large increase in relative risk in 2004. From 2005 onwards, there seems to be no obvious pattern in the data. The overall

average risk of NO_2 for the study period is 6%.

Table 4.3 shows the relative risk and the 95% confidence interval for the other three covariates of interest in the study. The results shown are from the models fitted with PM_{10} as the pollution covariate. The results for the models where NO_2 is the pollution covariate were similar, so are not shown here.

Table 4.3: Table of relative risks and 95% Confidence intervals for other covariates.

Year	Smoking	Log house price	Log Ethnic
2002	1.272,(1.241, 1.303)	0.947,(0.925, 0.969)	0.975,(0.957, 0.993)
2003	1.242,(1.212, 1.273)	0.914,(0.896, 0.937)	0.990,(0.972, 1.008)*
2004	1.242,(1.212, 1.273)	0.941,(0.920, 0.962)	0.964,(0.946, 0.981)
2005	1.241,(1.204, 1.269)	0.923,(0.902, 0.945)	0.974,(0.955, 0.994)
2006	1.218,(1.192, 1.245)	0.891,(0.870, 0.912)	0.971,(0.951, 0.991)
2007	1.230,(1.205, 1.256)	0.900,(0.877, 0.922)	0.985,(0.967, 1.004)*
2008	1.261,(1.237, 1.285)	0.903,(0.882, 0.924)	0.961,(0.942, 0.980)

Table 4.4: Table of ω increase for other covariates.

Year	Smoking	Log house price	Log Ethnic
ω	9.637	0.408	1.214

Smoking, not surprisingly, has the largest effect on respiratory disease. The relative risk of respiratory disease for an ω increase in smoking prevalence is between 21.8% and 27.2%, with an average of 24.4%.

Table 4.3 confirms that there is a significant negative linear relationship between log house price and log risk of respiratory disease in every year of the study; as house price increases the risk of respiratory disease decreases. The

smallest decrease in relative risk is in 2002, when an increase in log house price decreases the relative risk of respiratory disease by an average of 5.3%. In 2006 there is the largest relative risk decrease for log house price as the decrease in relative risk for this year is 10.9%. The overall average relative risk of log median house price is a 8.3% decrease.

Table 4.3 also shows a negative linear relationship between the natural log of the proportion of ethnic children and log relative risk of lung disease However, in 2003 and in 2007, the proportion of ethnic population is not significant in predicting admission to hospital with respiratory disease (as indicated by a *). Table 4.3 shows that for an ω increase in ethnic population, the largest decrease is 3.9% in 2008. The smallest significant decrease in relative risk for proportion of ethnic population is 2.5% in 2002. The overall decrease in relative risk for percentage of children from ethnic minorities is 2.6%.

We now need to check if these models are a good fit for the data. To do this we shall look at plots of residuals versus fitted values, as well as calculate the Moran's I and overdispersion statistics. As the residual plots all look very similar, we shall only look at the plots of the models from the start, middle and end of the study period, i.e 2002, 2005 and 2008.

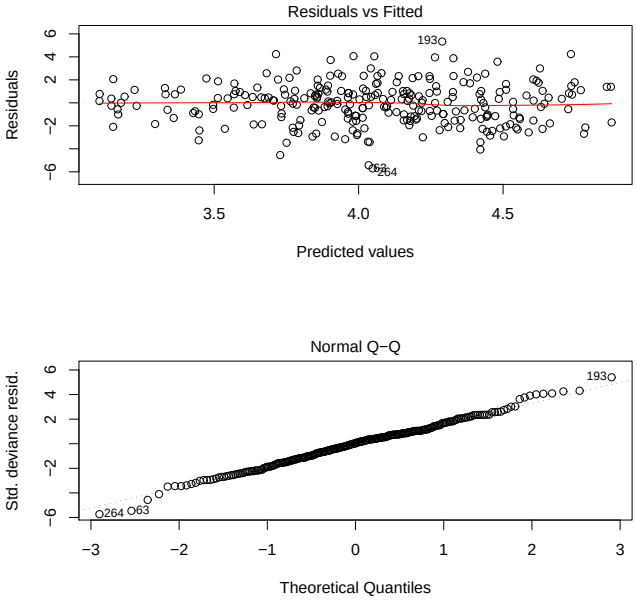

Figure 4.1: Residual vs fitted value and normal QQ plot for 2002 model on PM_{10}.

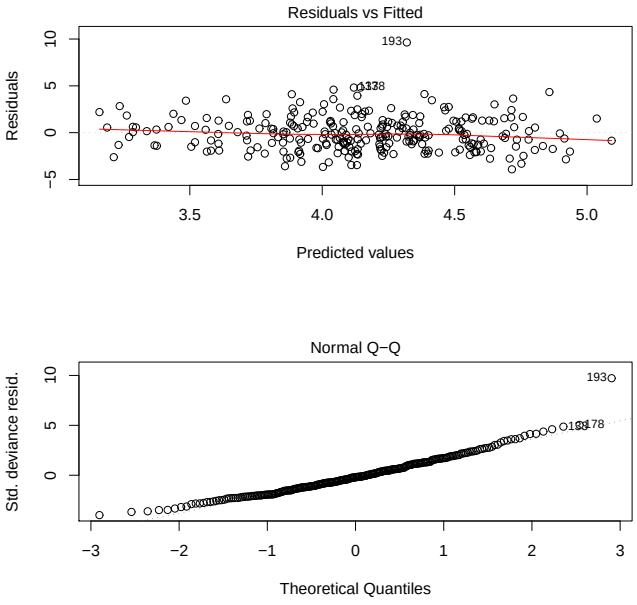

Figure 4.2: Residual vs fitted value and normal QQ plot for 2005 model on NO_2.

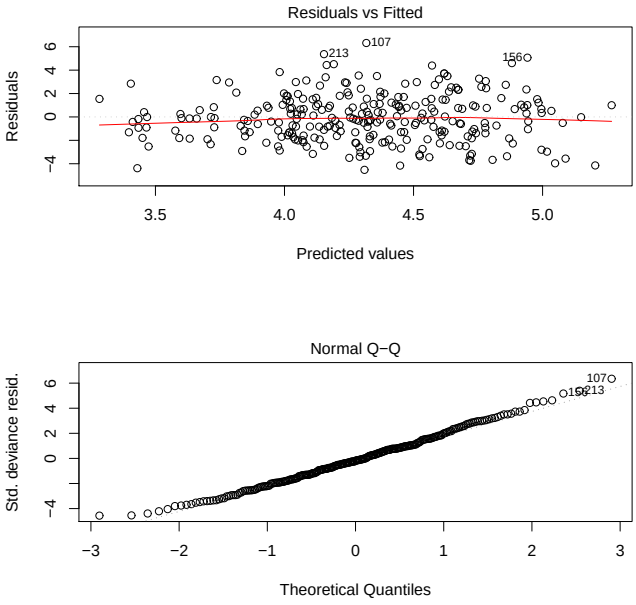

Figure 4.3: Residual vs fitted value and normal QQ plot for 2008 model on PM_{10}.

Figure 4.1 shows no obvious pattern in the plot of residuals versus fitted values, and the points seem evenly spread around zero. The normality plot in Figure 4.1 also follows the line of normality, satisfying the assumption of normality in the residuals. The plot of residual versus fitted values in Figure 4.2 also shows no obvious pattern and the points all seem evenly spread around zero. The normality plot in Figure 4.2 seems to follow the line of normality, with perhaps some slight deviation from the line at the tail ends, overall satisfying the assumption of normality. In Figure 4.3 there again seems to be no obvious pattern, and the QQ plot seems to follow the line of normality.

Table 4.5 shows the overdispersion statistics and the Moran's I statistic for the residuals of the models of PM_{10}. The results for the NO_2 models were similar and are not shown.

Table 4.5: Table of Moran's I and $\hat{\phi}$ for the residuals from each year.

Year	$\hat{\phi}$	**Moran's I**	**Moran's I p value**
2002	3.13	0.1565	0.0001
2003	3.67	0.1584	0.0001
2004	3.65	0.0898	0.0097
2005	3.65	0.0537	0.0617
2006	3.07	-0.0054	0.5124
2007	3.77	0.0022	0.4283
2008	4.22	0.0427	0.1057

Table 4.5 shows that the $\hat{\phi}$ statistic in every year is greater than 1, ranging between 3.07 and 4.22. Therefore we can conclude that mean and variance of the data are not equal as was assumed by the Poisson model, and there is overdispersion in every year of the study period. The models for 2002, 2003 and 2004 all show substantial spatial correlation within the residuals,

as the p values are all less than 0.05. Therefore we can conclude that the assumption of independence in the model is not satisfied for these years. The residuals for 2005 are not significant as the p value is 0.0617, which is just greater than 0.05. It could be argued that the residuals for 2005 model are borderline significant. The models from 2006-2008 have very small Moran's I statistics that are not significant, suggesting there is no evidence of spatial correlation in these residuals. Therefore we cannot reject the null hypothesis that the data are independent in these years of the study. However, there is evidence of spatial correlation when the Moran's I statistic is calculated for the SIR of each year. We calculated these Moran's I statistics to check if there is higher spatial correlation in the raw data from 2002 to 2004 compared with 2005 to 2008 which might explain the results in Table 4.5.

Table 4.6: Table of Moran's I for SIR.

Year	**Moran's I**	**Moran's I p value**
2002	0.4326	0.0001
2003	0.4193	0.0001
2004	0.3851	0.0001
2005	0.4082	0.0001
2006	0.4144	0.0001
2007	0.4262	0.0001
2008	0.4029	0.0001

Table 4.6 shows spatial correlation, on average, of 0.4127 throughout the study period. The lowest value of spatial correlation is 0.3851 in 2004, and the highest value is 0.4326 in 2002. Table 4.6 also shows the p values of Moran's I statistic are less than 0.05 for every year of the study. Unfortunately this does not show the same pattern as Table 4.5 before with strong spatial correlation from 2002 to 2004, then slightly weaker spatial correlation for the rest of the study period. The simple Poisson models in this

section have proved to be inappropriate for our data, as they do not allow for the overdispersion or spatial correlation in the data. We will now fit our Bayesian conditional autoregressive models, which will account for the spatial correlation and overdispersion in the data.

4.2 Conditional autoregressive models

We will now fit conditional autoregressive models to take into account any spatial correlation and overdispersion that exists within the data. However, we need to choose what kind of CAR model to fit from the methods discussed in Chapter 2. We found in the previous section that a range of spatial correlation exists within the residuals from the simple model, from strong spatial correlation, to very weak or no spatial correlation what so ever. As discussed in Chapter 2, the intrinsic CAR model is only appropriate for strong spatial correlation, so it is not really suitable for our data. The proper CAR model however allows for a range of correlation strengths, and if spatial correlation does not exist, the proper CAR model can represent independence. We shall therefore fit the proper CAR model given by equation (2.18) to each year of the study separately. First we shall look at the effect of air pollution on respiratory disease. Again we shall fit separate models for PM_{10} and NO_2, and include smoking prevalence, log median house price and percentage of children from ethnic minorities as additional covariates. The Bayesian model we are fitting is therefore

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k), \\ \ln(R_k) &= \mu + \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta} + \phi_k, \qquad (4.2) \\ f(\phi_k|\phi_{j\sim k}) &\sim N\left[\rho\frac{1}{n_k}\sum_{j\sim k}\phi_j, \frac{\tau^2}{n_k}\right] \qquad (4.3) \end{aligned}$$

where $f(\phi_k|\phi_{j\sim k})$ is the informative prior on the random effects that models the overdispersion and spatial correlation within the data. There are also non-informative priors on $\boldsymbol{\beta}$, μ, ρ, and τ^2. We have fitted $\boldsymbol{\beta} \sim \text{N}(0, 1\times10^6)$, as each β_i could be any real number. We have also fitted the same prior to μ. The value of ρ can be any value between 0 and 1 so we have fitted $\rho \sim \text{U}(0,1]$. As τ^2 controls the variance within ϕ_k, it has to be a positive number. We have therefore fitted $\tau^2 \sim \text{U}[0, 1000]$.

Table 4.7: Table of relative risks for PM_{10} for the proper CAR models.

Year	Relative risk	95% credible interval
2002	1.021	(0.9591, 1.082)*
2003	1.051	(0.9921, 1.123)*
2004	1.041	(0.9939, 1.089)*
2005	1.036	(0.9979, 1.077)*
2006	1.029	(0.9903, 1.070)*
2007	1.047	(1.0110, 1.083)
2008	1.022	(0.9844, 1.059)*

Table 4.8: Table of relative risks for NO_2 for the proper CAR models.

Year	Relative risk	95% credible interval
2002	1.027	(0.9703, 1.087)*
2003	1.058	(0.9959, 1.130)*
2004	1.050	(1.0020, 1.100)
2005	1.037	(0.9961, 1.079)*
2006	1.037	(0.9974, 1.076)*
2007	1.042	(1.0080, 1.079)
2008	1.028	(0.9918, 1.065)*

Tables 4.7 and 4.8 show that for most of the study period, PM_{10} and NO_2 are not significant predictors for relative risk of respiratory disease, as the 95% credible intervals contain 1. The only intervals that do not contain 1 are PM_{10} in 2007, and NO_2 in 2004 and 2007. Table 4.7 shows that for an ω increase in PM_{10} in 2007, the relative risk of respiratory disease increases by 4.7%, with an increase of anything between 1.1% and 8.3%. For NO_2 in 2004, the mean increase in relative risk for an increase in NO_2 of ω units is 5%, and the credible interval is between 0.2% and 10%. In 2007, the mean increase in relative risk is 4.2% with a credible interval of 0.8% and 7.9%. The overall average relative risk of PM_{10} throughout the study period is a

3.5% increase, and for NO_2 it is 4%.

Table 4.9: Table of relative risk for other covariates in PM_{10} models.

Year	Smoking	Log house price	Log Ethnic
2002	1.267,(1.209, 1.326)	0.951,(0.911 ,0.992)	0.966,(0.929, 1.006)*
2003	1.220,(1.161, 1.282)	0.892,(0.853, 0.934)	0.994,(0.954, 1.035)*
2004	1.254,(1.190, 1.320)	0.934,(0.889, 0.979)	0.980,(0.945, 1.018)*
2005	1.254,(1.120, 1.309)	0.915,(0.877, 0.954)	0.993,(0.958, 1.027)*
2006	1.253,(1.202, 1.306)	0.885,(0.850, 0.920)	0.966,(0.931, 1.002)*
2007	1.244,(1.191, 1.299)	0.901,(0.864, 0.934)	0.987,(0.955, 1.020)*
2008	1.277,(1.225, 1.332)	0.902,(0.865, 0.939)	0.975,(0.942, 1.010)*

Table 4.10: Table of relative risk for other covariates in NO_2 models.

Year	Smoking	Log house price	Log Ethnic
2002	1.264,(1.207, 1.324)	0.965,(0.926 ,1.003)*	0.951,(0.911, 0.993)
2003	1.213,(1.152, 1.278)	0.890,(0.850, 0.934)	0.992,(0.952, 1.032)*
2004	1.248,(1.184, 1.315)	0.933,(0.900, 0.979)	0.976,(0.939, 1.014)*
2005	1.252,(1.197, 1.309)	0.914,(0.876, 0.954)	0.991,(0.956, 1.027)*
2006	1.250,(1.200, 1.301)	0.884,(0.849, 0.919)	0.963,(0.929, 1.001)*
2007	1.244,(1.191, 1.297)	0.898,(0.860, 0.935)	0.988,(0.955, 1.022)*
2008	1.274,(1.222, 1.331)	0.901,(0.865, 0.939)	0.971,(0.939, 1.004)*

Table 4.9 shows the relative risks of smoking, log house price and log ethnic from the PM_{10} models and Table 4.10 shows the results of these covariates from the NO_2 models. There is a slight difference between the results of the two sets of models. Table 4.9 shows that both smoking and the log transformation of the median house price are significant in every year of the study as the 95% credible intervals do not contain 1, and our ethnic covariate is not significant in every year of the study as all the 95% credible intervals contain 1. Table 4.10 shows that from 2003 to 2008 that smoking and log

house price are again significant and log ethnic is not significant in the NO_2 models. However, in 2002 the log house price covariate is not significant in this model as the 95% credible interval contains 1, and conversely log ethnic is significant in this model as it is the only log ethnic 95% credible interval that does not contain 1.

Both Table 4.9 and Table 4.10 show again that smoking has the largest effect on relative risk, as for an ω increase (from Table 4.4) in prevalence of smokers there is an increase in relative risk of between 22.0% and 27.7% with an average increase of 25.3% throughout the study period for the PM_{10} models, and an increase between 21.3% and 27.4% for the NO_2 models with an average increase of 24.9%. There is a negative linear relationship between log house price and hospital admission risk with respiratory disease, that is as house price increases the risk of respiratory disease decreases. The smallest significant decrease in relative risk are 6.6% in the PM_{10} models and 6.7% in the NO_2 models, and the largest decreases are 11.5% and 11.6%. The average decrease for log median house price through all years of the study is 8.9% for the PM_{10} models, and 7.5% for the NO_2 models. The only year in the study that the population of ethnic children is a significant covariate is 2002 in Table 4.10. In 2002, there is a 4.9% mean decrease in relative risk of respiratory disease for an increase in the non-white population.

To investigate the strength of the underlying spatial correlation, we looked at the posterior distributions of the spatial correlation parameter ρ from equation (2.18) as shown in Table 4.11. As both tables for PM_{10} and NO_2 are similar, only the values for PM_{10} are shown.

Table 4.11 shows strong values of ρ for 2002 and 2003, and the credible intervals are close to the mean value. There is then a large drop in ρ between 2003 to 2004 from 0.851 to 0.496, and the credible intervals start to get a

Table 4.11: Table of ρ and σ^2 for proper CAR.

Year	ρ	τ^2
2002	0.847,(0.612, 0.975)	0.128,(0.088, 0.176)
2003	0.845,(0.616, 0.974)	0.156,(0.114, 0.207)
2004	0.489,(0.068, 0.841)	0.194,(0.146, 0.256)
2005	0.356,(0.028, 1.730)	0.172,(0.132, 0.221)
2006	0.345,(0.029, 0.704)	0.170,(0.131, 0.235)
2007	0.220,(0.009, 0.575)	0.176,(0.136, 0.223)
2008	0.278,(0.018, 0.632)	0.187,(0.145, 0.236)

lot wider here, suggesting that the spatial correlation is less strong compared with previous years. There is a decrease in ρ every year from 2004 to 2007, with a slight increase from 0.219 in 2007 to 0.271 in 2008. The values for ρ seem to follow a similar pattern as the Moran's I statistics from Table 4.5. Also shown in Table 4.11 is the posterior distribution of the variances τ^2 which increase from 0.128 in 2002 to 0.194 in 2004. From 2005 to 2008 the values increase and decrease between 0.172 and 0.187.

To investigate how well the chains converged, we can look at the history plot of the chains and check they overlap. Figure 5.1 shows the history plots for the relative risk of PM_{10} for 2002, 2005, and 2008. Figure 5.1 shows that for 2002, 2005 and 2008 the markov chains for the relative risk parameter for PM_{10} converges. The history plots for all the other covariates also showed this pattern.

To investigate that the CAR models had been fitted correctly and there is no leftover overdispersion and spatial correlation, we again calculated the overdispersion statistic and Moran's I statistic on the residuals of all the fitted models and ran permutation tests to check if Moran's I was significant. The p values of these test all came back greater than 0.05 and the overdispersion

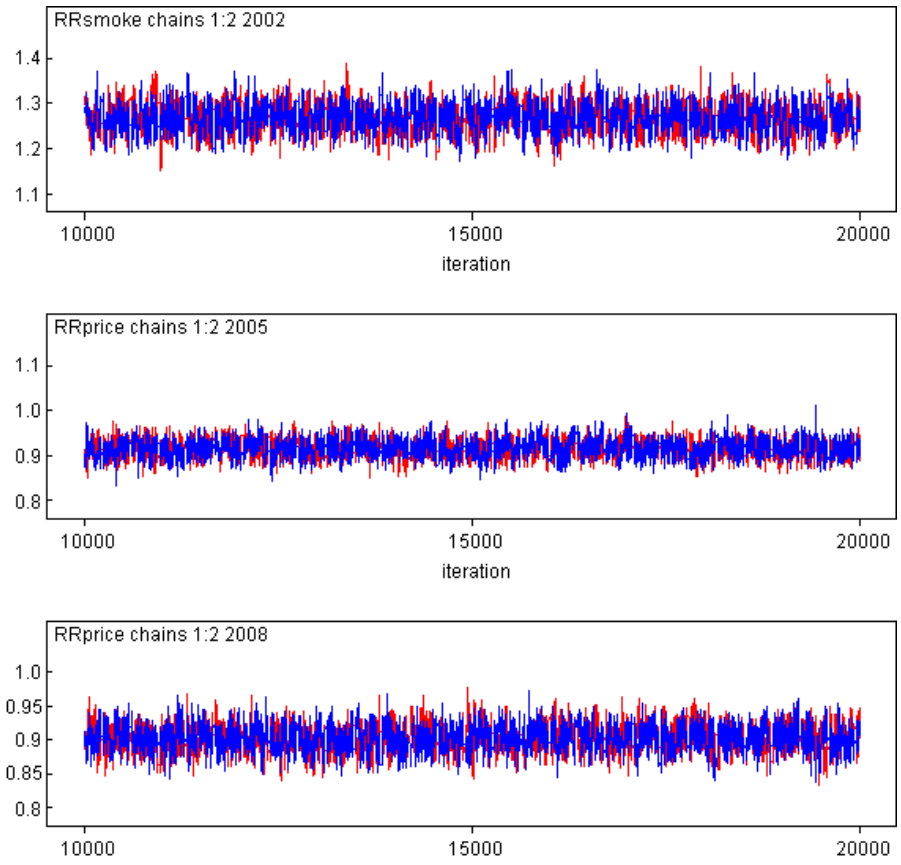

Figure 4.4: History Plot of chains for the Relative Risk of PM_{10}.

statistics were all less than 1. Therefore we can conclude that there is no longer any spatial correlation or overdispersion in the data and that the proper CAR model appears to be a good fit for the data.

Chapter 5

Spatial-temporal health models

We are now going to look at the effects of air pollution and the other covariates on respiratory disease through time as well as in space. We now fit two models to the health data, one with PM_{10} and one with NO_2 as the exposure of interest. The model we are initially fitting has no correlation terms and is given by

$$\begin{aligned} Y_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\ \ln(R_{kt}) &= \mu + \mathbf{x}_{kt}^{\mathrm{T}}\boldsymbol{\beta}, \end{aligned} \tag{5.1}$$

where t=1,...,7 denotes the year of the study. The relative risks associated with a one standard deviation increase in each covariate are shown in Table 5.1. As the covariate effects are very similar using either exposure, the results for the PM_{10} model are shown. The smoking covariate only has information for one year and hence the same set of values are used for each of the seven years. In addition, we also used the 2004 ethnic variable in the years 2002 and 2003 as it was measured in those years.

Table 5.1 shows that throughout the study period, PM_{10} has a mean relative risk increase of 4.6% in respiratory disease cases for an increase of 1.986. This relative risk is similar than the overall average PM_{10} increase for all the

Table 5.1: Table of relative risks for space time models.

Year	**Relative risk**	**95% confidence interval**
PM_{10}	1.046	(1.040, 1.053)
NO_2	1.049	(1.042, 1.056)
smoking	1.242	(1.231, 1.253)
ethnic	0.980	(0.973, 0.984)
house price	0.914	(0.906, 0.922)

individual relative risks in the proper CAR model discussed in Chapter 4, which was 3.5%. Table 5.1 also shows that the 95% confidence interval does not contain 1, so PM_{10} has a significant effect on respiratory disease cases in Greater Glasgow and Clyde across the entire study period. The mean relative risk increase for NO_2 in Table 5.1 is 4.9%, and again the confidence interval does not contain 1 so NO_2 is significant in predicting respiratory disease cases. This figure is similar to the overall average increase of all seven yearly NO_2 values calculated in Table 4.8, which was a 4.0% increase. For our other three covariates of interest, we see that they are all significant in predicting the risk of respiratory disease, as none of the credible intervals contain 1. Smoking again has the largest effect on respiratory disease hospitalisation, with a 24.2% increase in cases for an increase of 9.637% in smoking prevalence. There is again negative linear relationships for the proportion of ethnic children and the median house price variables with respiratory risk. For a 1.214% increase in the proportion of non-white children within each small area there is a 2% decrease in respiratory disease cases thought the study period and area, and there is an 8.6% decrease in respiratory disease cases for an increase in median house price.

To investigate how well the models were fitted, we can look at plots of residuals versus fitted values and plots of normality for each model. Figure 5.1 shows the plots of the PM_{10} model, and Figure 5.2 shows the plots of the

model for NO_2.

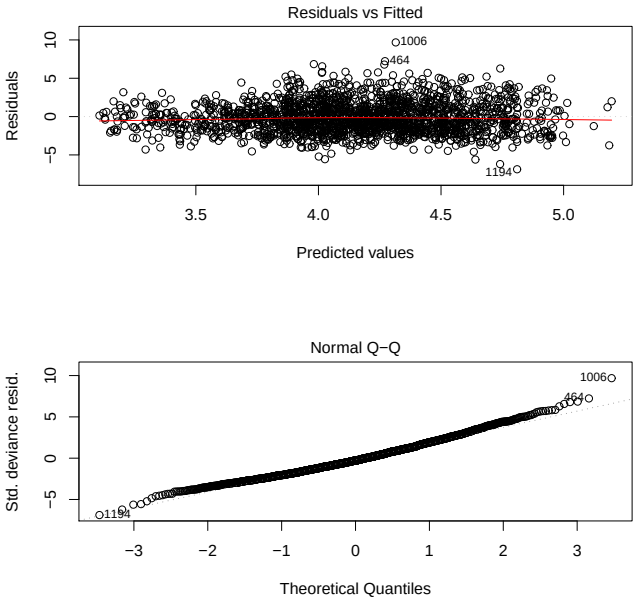

Figure 5.1: Residuals vs fitted and normal QQ plot of PM_{10}.

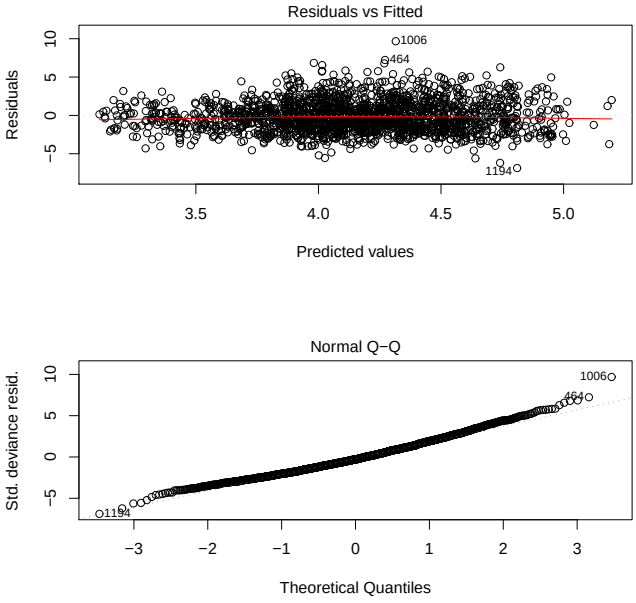

Figure 5.2: Residuals vs fitted and normal QQ plot of NO_2.

Both Figures 5.1 and 5.2 show no obvious pattern in the plot of residuals

versus fitted values, and the points all seem evenly spread around zero. The plot of normality shows that both models follow the line of normality, satisfying the assumption of normality in the residuals.

From the previous model we have found that all our covariates of interest are significant at predicting admission to hospital with respiratory disease throughout the study period. However, we have to check for spatial correlation by calculating Moran's I statistic and also the level of overdispersion statistic. We calculated the Moran's I statistic as shown in (2.9), for each year individually by splitting the residuals into the appropriate group. The Moran's I statistic as well as the p values are shown in Table 5.2. The results for the PM_{10} model and NO_2 model are very similar so the results for the PM_{10} are shown.

Table 5.2: Table of Moran's I for space and time model.

Year	**Moran's I**	**Moran's I p value**
2002	0.0735	0.021
2003	0.0134	0.316
2004	-0.0085	0.535
2005	0.046	0.091
2006	0.0413	0.116
2007	0.0953	0.005
2008	0.0769	0.017

The overdispersion statistic we calculated gave us a value of 4.07 which tells us there was significant overdispersion. Table 5.2 shows there is only significant spatial correlation within the data in 2002, 2007 and 2008, with 2005 being borderline as it is just above the 0.05 significance level. The Moran's I statistics for the years of significant spatial correlation are 0.0735 in 2002, 0.0953 in 2007 and 0.0769 in 2008. Though this is very weak spatial

correlation, we have shown that it is significant in the data and will have to be accounted for within the model. Table 5.2 also shows that the p values for 2003 to 2004, and 2006 are all much larger than 0.05, so there is no significant spatial correlation for these years and the data can be assumed to be independent. We can now add in the random effects that will model the spatial correlation and overdispersion in the data as well as any temporal effects that exist. The model we fit is based on the model proposed by Knorr-Held (2000), where instead of two spatial priors, one that assumes strong spatial correlation and one that assumes independence in the data, we shall fit the proper CAR model outlined in (2.18) which allows for a range of spatial correlation as well as independence, which we have shown exists within our model. The model we fit to the data is now

$$\begin{aligned} Y_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\ \ln(R_{kt}) &= \mu + \mathbf{x}_{kt}^{\mathrm{T}}\boldsymbol{\beta} + \alpha_t + \phi_k, \\ f(\alpha_t) &\sim N\left(\alpha_{t-1}, \tau_\alpha^2\right) \\ f(\phi_k|\phi_{j\sim k}) &\sim N\left[\rho\frac{1}{n_k}\sum_{j\sim k}\phi_j, \frac{\tau_\phi^2}{n_k}\right] \end{aligned} \tag{5.2}$$

where $f(\alpha_t)$ is the informative prior that models the temporal correlation via a first order random walk, and $f(\phi_k|\phi_{j\sim k})$ is the proper CAR prior that models the spatial correlation as before. We use the following non-informative priors on the remaining coefficients in the separable model; $\boldsymbol{\beta_i} \sim \text{N}(0, 1\times10^6)$, $\mu \sim \text{N}(0, 1 \times 10^6)$, $\rho \sim \text{U}[0, 1)$, $\tau_\phi^2 \sim \text{U}[0, 1000]$ and $\tau_\alpha^2 \sim \text{U}[0, 1000]$. The relative risks for the covariates of the model are shown in Table 5.3.

Table 5.3 shows similar results to those in Table 5.1. All covariates again are significant as none of the 95% Credible intervals contain 1. The relative risk increase of respiratory disease cases for PM_{10} is 4.1%, and for NO_2 the relative risk increase is 4.2%. For an increase in smoking prevalence, there is a 23.7% of respiratory disease cases in each small area through the seven

Table 5.3: Table of relative risks for space time model.

Year	**Relative risk**	**95% credible interval**
PM_{10}	1.041	(1.031, 1.051)
NO_2	1.042	(1.032, 1.053)
smoking	1.237	(1.225, 1.249)
ethnic	0.980	(0.973, 0.988)
house price	0.912	(0.903, 0.920)

year study period. For the ethnic covariate and the median house price covariate, there is a 2.0% and a 8.8% decrease respectively in mean relative risk.

To investigate the strength of the underlying spatial correlation and temporal correlation, we looked at the posterior distributions of the spatial correlation parameter ρ from equation (5.2) as shown in Table 5.4.

Table 5.4: Table of ρ and τ^2 for proper CAR.

Pollution model	ρ	τ_ϕ^2	τ_α^2
PM_{10}	0.806,(0.605, 0.945)	0.044,(0.041, 0.050)	0.004,(0.002, 0.018)
NO_2	0.821,(0.628, 0.951)	0.043,(0.034, 0.054)	0.004,(0.002, 0.018)

Table 5.4 shows a strong value of ρ for both the PM_{10} and NO_2 models of 0.806 and 0.821 respectively, suggesting there is strong underlying spatial correlation in the data. We also see that both the spatial variance, τ_ϕ^2 and the temporal variance τ_α are very small. The values of τ_α appear to be the same for the credible interval for τ_α^2 is very narrow, ranging from 0.041 to 0.050. To again check how well the models have converged, we look at the history plot of the chains and check if they overlap. Figure 5.3 shows the history plots for the two relative risk of our pollution covariates as before. Figure 5.3 shows that the markov chains for the relative risk parameter of

both parameters converges once again.

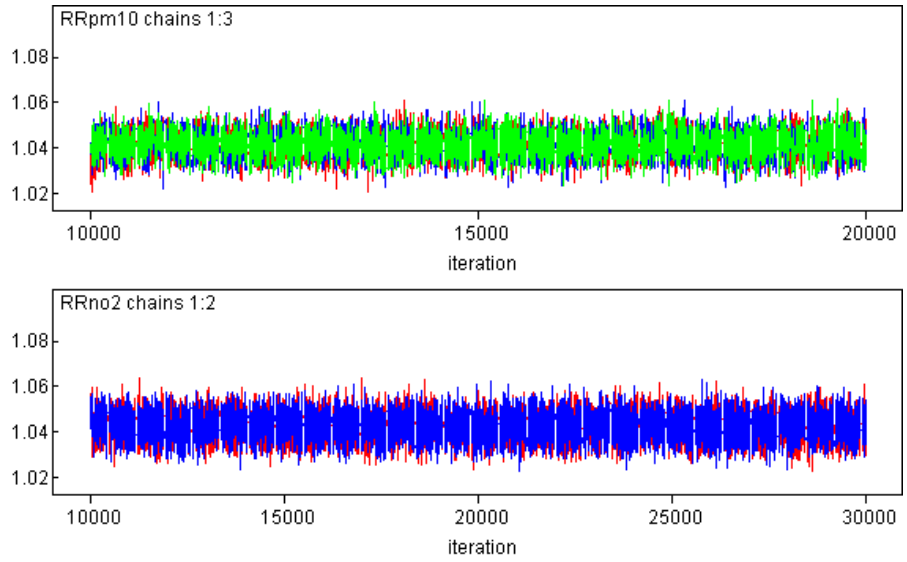

Figure 5.3: History Plot of chains for the Relative Risk of PM_{10}.

To check the proper CAR model has modeled underlying spatial correlation we calculated Moran's I for the residuals of the separable models for NO_2 and PM_{10}. There was no underlying spatial correlation as all the p values were greater than 0.05. We can therefore conclude that our model appears to be a good fit for the data.

Chapter 6

Conclusions

The main aim of this thesis was to investigate the effect that air pollution concentrations have on the number of respiratory disease hospital admissions in the Greater Glasgow and Clyde NHS health board. The data used were acquired from the Scottish Neighbourhood Statistics website (run by the Scottish Government), and the Department for the Environment, Food and Rural Affairs (DEFRA). The health response was respiratory disease hospitalisation, including patients admitted to hospital on a primary diagnosis of respiratory diseases apart from lung cancers. The covariates of interest included were smoking prevalence, median house price (a measurement of socio-economic deprivation), and the percentage of non-white school children in each area, which was chosen to represent the potential effects of ethnicity. The pollution measurements included were particulate matter less than 10 μgm^{-3}(PM_{10}) and nitrogen dioxide (NO_2) as modelled concentrations, which were available from DEFRA at the 1 km resolution. The modelled estimates were transformed to the IG scale by calculating the median value within each IG, and calculating the closest values for those IGs that were too small to contain a single grid square.

6.1 Results of the study

The modelling began with overdispersed Poisson log-linear models applied to all the covariates separately for each year with the exception that either PM_{10} or NO_2, but not both were included in a model. The over dispersion parameters for both the PM_{10} and NO_2 models were greater than 1, showing evidence of overdispersion in the data. A permutation test based on Moran's I statistic applied to the residuals found significant evidence of spatial correlation between 2002 and 2004. However, from the later years of the study, the results from the Moran's I statistics shows evidence that the spatial correlation in these models was not significant, ie the residuals in these models were independent. Therefore the models were extended to allow for overdispersion and spatial correlation by adding random effects to the model. This was achieved using conditional autoregressive (CAR) models.

Having fitted the models separately for each year with PM_{10} as the pollution covariate, it was only significant in 2007. In contrast, when NO_2 was included it exhibited a significant relative risk in 2004 and 2007. This is in contrast with the results from the models without random effects which exhibited significant results in every year. The most likely reason for these differences is that the Bayesian models correctly allow for additional uncertainty via the inclusion of the random effects, thus increasing the uncertainty intervals for the regression parameters. Thus from a purely spatial analysis we conclude that both pollutants exhibited substantial effects on the risk of hospital admission with respiratory disease in 2007, while for NO_2 it was also significant in 2004. Of the other three covariates, smoking prevalence was the only covariate significant in every year of the study for both pollutants. In addition, the only year that the log house price covariate was not significant in predicting respiratory disease was in the 2002 NO_2 model, and conversely

the only year that the log ethnic population covariate was significant was in the 2002 NO_2 model. To investigate that the underlying spatial correlation and overdispersion in the data had been accounted for by the proper CAR model Moran's I and overdispersion statistics were computed for the residuals. The overdispersion statistics were all close to 1 and the Moran's I statistics were all non-significant. This suggests that the proper CAR model is a good fit for the data as there is no longer any underlying spatial correlation or overdispersion within the data.

Finally, a spatio-temporal model was fitted that modelled the data from multiple years simultaneously. The temporal correlation was modelled via a first order random walk, with the spatial correlation again modelled by a proper CAR model. In these spatio-temporal analyses both pollutants exhibited significant health effects, with relative risks of 1.041 and 1.042 for a $1.986\mu gm^{-3}$ and a $7.032\mu gm^{-3}$ increase in PM_{10} and NO_2 respectively. The results of the space-time differ slightly from the spatial proper CAR models, as NO_2 was only significant in 2004 and both PM_{10} and NO_2 were significant in 2007. As before the presence of overdispersion and spatial correlation in the residuals were assessed, and no evidence of either was found. As the spatio-temporal models fit all the yearly data in one model per pollution covariate, instead of seven individual yearly models then the spatio-temporal model has seven times the data and hence will have more precise estimates and narrower uncertainty intervals. The spatio-temporal model is therefore the better, more accurate model and is the best model for fitting and analysing the data in this study.

Overall from the spatio-temporal analysis (which is based on the largest volume of data) we can conclude that long term exposure to both PM_{10} and NO_2 have a significant effect on admission to hospital with a primary diagnosis of respiratory disease in the Greater Glasgow and Clyde NHS healthboard

during the study period of 2002 to 2008. These results are concurrent with the results found in previous air pollution studies including Lee et al. (2009) and Elliot et al. (2007). In Lee et al. (2009) it was found that for a $1.7\mu gm^{-3}$ and $8\mu gm^{-3}$ increase in PM_{10} and NO_2 respectively, there was a relative risk increase of 1.07 and 1.09 in Greater Glasgow and Clyde. These results are slightly larger that the relative risks found in this study, but overall show that both pollutants have a negative effect on respiratory disease over a long term exposure in Greater Glasgow and Clyde. While Elliot et al. (2007) was looking at the long term exposure effect of different pollutants on a mortality of multiple health data including respiratory disease, the results showed that long term exposure to pollutants had the largest negative effect on respiratory disease.

6.2 Discussion

One problem with the data is that the pollution data was based on modelled estimates, and are not true values of pollution levels. As there are so few pollution monitoring stations within the study region the modelled pollution estimates are the only practical approach. To properly analyse these data, an error term should be added to the models.

Another problem with the data was the lack of smoking prevalence data for each year. It is safe to assume that smoking prevalence will change from year to year due to natural variation and that using only one year of data for every year of the analysis is not adequate. One possible cause of this variation is the Scottish public smoking ban of March 2006, which outlawed smoking in indoor public areas such as bars and clubs, restaurants, public transport and stations, and workplaces. This will have hit the number of smokers as it will have created a very large incentive or inspiration for some people to try quit. The global recession of 2008 may also have had an impact on smoking prevalence, as people were tightening their belts financially and cigarettes will be a luxury that people may have wanted to save money on. However, an alternative viewpoint is that the recession may have increased smoking prevalence as people who were made redundant may have started smoking more to deal with the stress of financial worries and unemployment. Thus as the smoking variable is based on data from 2001, 2003 and 2004, it may not give the best overall interpretation of the smoking prevalence for the entire study period.

A major statistical drawback of the study design is that it is an ecological study, and thus the pollution effect estimates used are based on population rather than individual level data. Therefore these results can only represent the effects of air pollution increases on the population on the whole, as op-

posed to for each individual. Assuming individual ecological estimates are the same is the ecological fallacy (Schwartz (1994)). Despite an individual study design being the preferable study design, the practicality of carrying out this kind of study is mostly impossible as you would require each person within your study region to carry a pollution monitor which is both impractical and expensive.

Ecological studies like this one are not un-important. The benefits of studies such as these are that they are quick and inexpensive to perform, allowing hypotheses to be generated about potential exposures of interest. Furthermore, as the results are based on population as a whole, ecological studies are informative for groups who are concerned with safeguarding overall public health. The results of an ecological study like this can show quickly the rapid deterioration of the public health due to increased exposure which in turn can lead to rapid action.

Another drawback of this study is only using the previous years air pollution data because it ignores exposure before this time. To improve this study design, the average pollution concentrations over the previous two or three years could be included in each model, as opposed to the one year previous to the respiratorty data.

Although unlikely to have changed very largely between 2002 and 2004, it would have been more accurate and more productive to have the proportion of non-white school children within each IG for each year of the study. The proportion of non-white school children may also not be the best indicator of the ethnic population within a small area, as it does not take into account families with no children, families with children who do not go to school, mixed race families and families whose children go to a school outwith the area they live in. The indicator we chose to represent deprivation, median

house price within an area, was the most appropriate as it had the lowest correlation value with smoking prevalence than the other available covariates, thus minimising the possibility of collinearity. However, it is still highly correlated with smoking (-0.7049481). Preferably we would want another variable that represents deprivation within each small area that is not as related to smoking, however it has been shown that poorer people have poorer health and are much more likely to smoke, so there may be no deprivation indicator that does not have high correlation with smoking and we could have left it out completely.

If there was more time, I would have considered allowing for measurement error within the modelled pollution concentrations as these are modelled rather than true concentrations. It would also have been interesting to fit a spatio-temporal model that allows for an interaction between both the spacial and temporal correlation, a so called non-separable model. This would have allowed the level of spatial correlation to vary from year to year, which was evident from our data. It would also have been interesting to investigate the effect of other pollution covariates, to determined what effect they have on respiratory health, such as sulphur dioxide (SO_2) or carbon dioxide (CO_2). It would also have been interesting to investigate the effects of pollution on different diseases and in different health boards, both urban and more rural, and compare the results. For example one might compare the Lothian health board to Greater Glasgow, to compare the relative risk of respiratory disease between Glasgow and Edinburgh. Alternatively the Ayrshire and Arran health board would make an interesting comparison, comparing the urban areas of Glasgow and Paisley to a study area which has a more rural population, as well as a lot of towns and villages along the west coast of Scotland where there should be more fresh air coming in from the sea. It would also be interesting to look at data from later years, such as 2009 and 2010 and investigate whether the UK governments car scrappage scheme,

where they offered a £2000 discount on any brand new car for a trade in of any car over ten years old would have had an effect on pollution variables, particulary on IG's with large traffic congestions like Glasgow City Center or areas of the M8 and M74.

Bibliography

Banerjee, s., Carlin, B. & Gelfand, A. (2004), *Hierarchical modeling and analysis for spatial data*, illustrated edn, CRC Press.

Bell, M., Davis, D. & Fletcher, T. (2004), 'A Retrospective Assessment of Mortality from the London Smog Episode of 1952: The Role of Influenza and Pollution', *Environmental Health Perspectives* **112**(1), 6–8.

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

Carder, M., McNamee, R., Beverland, I., Elton, R., Tongeren, M., Cohen, G., Boyd, J. & Agius, R. (2008), 'Interaction effects of particulate pollution and cold temperature on cardiorespiratory mortality in Scotland', *Occupational and Enviromental Medicine* **65**, 197–207.

Clean Air Act (1956), HM Stationary Office.

Cressie, N. (1993), *Statistics for Spatial Data*, revised edn, Wiley: New York.

Davis, D., Bell, M. & Fletcher, T. (2002), 'A Look Back at the London Smog of 1952 and the Half Century Since', *Environmental Health Perspectives* **110**(12), A734–735.

Dominici, F., Daniels, M., Zeger, S. & Samet, J. (2002), 'Air pollution and motality; Estimating regional and national dose-response relationships', *Journal of the American Statistical Association* **97**, 100–111.

Elliot, P., Shaddick, G., Wakefield, J., de Hoogh, C. & Briggs, D. (2007), 'Long-term associations of outdoor air pollution with mortality in Great Britain', *Thorax* **62**, 10881094.

Fairbarn, A. & Reid, D. (1958), 'Air pollution and other local factors in respiratory disease', *British Journal of Preventive and Social Medicine* **12**, 94–103.

Gauderman, W., Avol, E., Lurman, F., Kuenzli, N., Gilliland, F., Peters, J. & McConnel, R. (2005), 'Childhood Asthma and Exposure to Traffic and Nitrogen Dioxide', *Epidemiology* **16**(6).

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

Haining, R., Li, G., Maheswaran, R., Blangiardo, M., Law, J., Best, N. & Richardson, S. (2010), 'Inference from ecological models: estimating the relative risk of stroke from air pollution exposure using small are data', *Thorax* **62**, 1088–1094.

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

Ismaila, A., Canty, A. & Thabane, L. (2007), 'Comparison of Bayesian and frequentist approaches in modelling risk of preterm birth near the sydney Tar Ponds, Nova Scotia, Canada', *BMC Medical Research Methodology* **7**.

Jerret, M., Buzzelli, M., Burnett, R. & DeLuca, P. (2005), 'Particulate air pollution, social confounders, and mortality in small areas of an industrial city', *Social Science and Medicine* **60**, 2845–2863.

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

Laden, F., Neas, L., Dockery, D. & Schwartz, J. (2000), 'Association of Fine Particulate Matter from Different Sources with Daily Mortality in Six U.S. Cities', *Environmental Health Perspectives* **108**, 941–947.

Lawson, A. (2009), *Bayesian Disease Mapping: Hierarchical modeling in spatial epidemiology*, illustrated edn, CRC Press.

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

Lee, D., Ferguson, C. & Mitchel, R. (2009), 'Air pollution and health in scotland: a multicity study', *Biostatistics* **10**, 409–423.

Longhurst, J., Irwin, J., Chatterton, T., Hayes, E., Leksmono, N. & Symons, J. (2009), 'The development of effects-based air quality management regimes', *Atmospheric Environment* **43**, 64–78.

Maheswaran, R., Haining, R., Brindley, P., Law, J., Pearson, T., Fryers, P., Wise, S. & Campbell, M. (2005), 'Outdoor air pollution and stroke in Sheffiled, United Kingdom.', *Stroke* **36**, 239–243.

Maheswaran, R., Haining, R., Pearson, T., Law, J., Brindley, P. & Best, N. (2006), 'Outdoor NO_x and stroke mortality: adjusting for small are level smoking prevalence using a Bayesian approach', *Statistical Methods in Medical Research* **15**, 499–516.

McColl, J. (1995), *Probability*, reprinted edn, Butterworth-Heinemann: Oxford.

Moran, P. (1950), 'Notes on Continuous Stochastic Phenomena', *Biometrika* **37**, 17–23.

Pandit, S. & Wu, S. (1983), *Time series and system analysis with applications*, illustrated edn, John Wiley and Sons Inc.

Parliamentary Office of Science and Technology (2000), 'Air quality in the uk', *Parlimentary Office of Science and Technology* **188**, 1–4.

Prescott, G., Cohen, G., Elton, R. & Agius, R. (1998), 'Urban air pollution and cardiopulmonary ill health: a 14.5 year time series study', *Occupational and Enviromental Medicine* **55**, 697–704.

Samoli, E., Schwartz, J., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F., Medina, S., Ross, G., Sunyer, J., Bacharova, L., Anderson, H. & Katsouyanni, K. (2001), 'Investigating regional differences in short-term effects of aitr pollution on daily mortality in the APHEA project; A sensitvity analysis for controlling long-term trends and seasonality', *Enviromental Health Perspectives* **109**(4).

Schwartz, S. (1994), 'The Fallacy of the Ecological Fallacy: The Potential Misuse of a Concept and the Consequences', *American Journal of Public Health* **84**(5).

Smith, G., Williams, F. & Lloyd, O. (1987), 'Respiratory cancer and air pollution from iron foundries in a Scottish town: an epidemiological and enviromental study', *British Journal of Industrial Medicine* **44**, 795–802.

Young, L., Gotway, C., Yang, J., Kearney, G. & DuClos, C. (2009), 'Linking health and enviromental data in geographical analysis: It's so much more than centroids', *Spatial and Spatio-temporal Epidemiology* **1**, 73–84.