Examining the effect of residual spatial autocorrelation on fixed effect estimation

Katie Jayne Stewart

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

School of Mathematics & Statistics

January 2015

Abstract

Estimation of fixed effects in spatial data sets can be challenging, as spatial autocorrelation can occur in the residuals as well as the covariates. The residual spatial autocorrelation can be caused by spatially autocorrelated risk factors for the response data that are unknown or unmeasured, and leads to unmeasured confounding. Spatial regression models have been developed to allow fixed effect estimation whilst accounting for residual spatial autocorrelation, and three of these methods have been compared here through a simulation study along with a method which ignores the spatial autocorrelation. The aim of this thesis is thus to determine if accounting for the spatial autocorrelation produces better results in terms of fixed effect estimation, and if so which method is the best. These aims are first examined through simulation studies, and then the methods are applied to a study of air pollution and respiratory illness hospital admissions in the central belt of Scotland in 2010. The analysis shows that higher concentrations of particulate matter air pollution result in an increased risk of hospital admission due to respiratory illness.

Acknowledgements

Thank you to everyone who has helped me throughout this year. In particular, Duncan Lee for all his help and guidance as my supervisor. Also my parents for their continued love and support.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated.

Contents

1 Introduction. **1**

2 Methods and Existing Literature. **6**

2.1 Data 6

2.2 Poisson Generalised Linear Model and Quasi-Likelihood 7

2.3 Moran's I 8

2.4 Bayesian Methods 9

2.5 Conditional Autoregressive (CAR) Models 11

2.6 Sparse Spatial Generalised Linear Mixed Model (SGLMM) . . 13

2.7 Localized Conditional Autoregressive (LCAR) Model 14

2.8 Model Performance 17

2.9 Existing Areal Unit Studies on Air Pollution and Health . . . 18

3 Investigating The Effects of Ignoring Residual Spatial Autocorrelation on Fixed Effect Estimates. **23**

3.1 Introduction 23

3.2 Data Generation 24

3.3 Results 28

3.3.1 Scenario 1 - Varying correlation in the residual structure with an uncorrelated covariate 28

3.3.2 Scenario 2 - Varying autocorrelation in the residual structure with a moderately autocorrelated covariate . 33

3.3.3 Scenario 3 - Varying autocorrelation in the residual structure with a strongly autocorrelated covariate . . . 38
3.3.4 Scenario 4 - Varying autocorrelation in the covariate with weakly autocorrelated residual structure 44
3.3.5 Scenario 5 - Varying autocorrelation in the covariate with moderately autocorrelated residual structure . . . 47
3.3.6 Scenario 6 - Varying autocorrelation in the covariate with strongly autocorrelated residual structure 50
3.4 Conclusion 52

4 Comparing The Effects of Ignoring and Accounting For Residual Spatial Autocorrelation on Fixed Effect Estimates. 54
4.1 Introduction 54
4.2 Data Generation and Study Design 55
4.3 Results 56
4.3.1 Small Impact Random Effects 57
4.3.2 Large Impact Random Effects 58
4.4 Conclusion 61

5 An Application to Central Belt Respiratory Health Data. 64
5.1 Introduction 64
5.2 Data 64
5.3 Modelling 74
5.4 Results 77
5.5 Conclusions 79

6 Conclusions. 81

List of Tables

4.1 RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 100 expected disease cases and Var$[\phi] = 0.1$. 59

4.2 RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 10 expected disease cases and Var$[\phi] = 0.1$. . 60

4.3 RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 100 expected disease cases and Var$[\phi] = 1$. . 60

4.4 RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 10 expected disease cases and Var$[\phi] = 1$. . . 61

5.1 Overdispersion Parameter for Quasi-Poisson GLMs. 74

5.2 Moran's I for Quasi-Poisson GLMs. 76

5.3 Relative Risk (95% Uncertainty Interval) for Pollutants Using Quasi-Poisson, Leroux, Sparse SGLMM and LCAR Models. . 77

5.4 DIC for each model. 78

List of Figures

3.1 RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases, and the expected number of disease cases is equal to 100. 31

3.2 RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1. 31

3.3 RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01. 32

3.4 RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10. 32

3.5 RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000. 33

3.6 RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases, and the expected number of disease cases is equal to 100. 36
3.7 RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1. 36
3.8 RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01. 37
3.9 RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10. 37
3.10 RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000. 38
3.11 RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 100. 41

3.12 RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1. 42
3.13 RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01. 42
3.14 RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10. 43
3.15 RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000. 43
3.16 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 100. . . 45
3.17 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 10. . . 46
3.18 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 1000. . 46
3.19 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of disease cases equal to 100. 48

3.20 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of diseases cases equal to 10. 49
3.21 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of disease cases equal to 1000. 49
3.22 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure and an expected number of disease cases equal to 100. . . 51
3.23 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure and an expected number of disease cases equal to 10. . . 51
3.24 RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure with an expected number of disease cases equal to 1000. . 52

5.1 Local Authority Map of Scotland. Source: The Scottish Government (http://www.scotland.gov.uk/Resource/Doc/933/0009386.pdf). 65
5.2 Map of the SIR for the Central Belt (2010). 67
5.3 Histogram of the SIR for the Central Belt (2010). 67
5.4 Boxplot of the Pollutant Concentrations (μgm^{-3}, 2009). 68
5.5 Map of the NO_2 (μgm^{-3}) for the Central Belt (2009). 69
5.6 Map of the SO_2 (μgm^{-3}) for the Central Belt (2009). 69
5.7 Map of the PM_{10} (μgm^{-3}) for the Central Belt (2009). 70
5.8 Map of the $PM_{2.5}$ (μgm^{-3}) for the Central Belt (2009). 70
5.9 Correlation between the four pollutants. 71
5.10 Map of Income Deprivation for the Central Belt (2009). 72
5.11 Map of Ethnicity for the Central Belt (2009). 72
5.12 Map of Urban for the Central Belt (2008). 73
5.13 Residual Plots for Quasi-Poisson GLM Including PM_{10}. 75

Chapter 1

Introduction.

Many data sets consist of measurements of the same underlying quantity at different locations in space, including measurements of population health such as disease risk and environmental exposures such as air pollution. The former typically relate to a set of non-overlapping contiguous areal units, such as the set of electoral wards in a city or county. Common aims in modeling spatial health data include estimating the geographical pattern in disease risk or estimating the impact of an exposure on disease risk. The latter is achieved using regression techniques, utilizing the spatial contrasts in both the exposure and the response.

However, the residuals from fitting a regression model to the disease and exposure data will typically contain spatial autocorrelation, which occurs when observations close together in space are likely to be more similar than those further apart. This spatial autocorrelation can be caused by numerous factors affecting disease risk, which themselves have a spatial pattern but have not been put into the regression model as covariates. When such factors are known and measured they can be included in the regression model, but if not then this leads to unmeasured confounding causing residual spatial autocorrelation. This unmeasured confounding has led to spatial regression models being developed, that can estimate the effects of covariates on disease

risk whilst accounting for residual spatial autocorrelation.

However, recent research (Clayton et al. (1993), Reich et al. (2006), Hughes and Haran (2013)) has shown that if a covariate in the model is spatially autocorrelated then its estimated regression coefficient can be biased, due to potential collinearity with the term in the model that allows for the residual spatial autocorrelation. This calls into question the usefulness of such spatial autocorrelation regression models, although no widescale study of this phenomenon has been undertaken. It also raises the questions: (i) does allowing for this residual autocorrelation produce better results in terms of covariate estimation than ignoring it; and (ii) if allowing for the autocorrelation is better, which is the best modelling approach to allow for it.

One field where this problem arises is studies investigating the effects of air pollution on human health, with examples being Dominici et al. (2006) and Lee et al. (2014). In this study design the data are the form of counts of occurences of the health outcome, and the focus of this thesis is on areal unit studies. Areal unit studies partition the region into n distinct regions that could be electoral wards or similar. The number of disease cases in each unit is then regressed against air pollution concentrations and other covariates in that unit, yielding a population level association. Pollution concentration data can be of two forms, either point level data or modelled data. Point level data comes from monitoring stations situated across the United Kingdom, with 9 monitoring stations in central Scotland. This provides exact measurements at the monitoring location, however, these do not cover all the intermediate geographies and there are more in cities where pollution is expected to be higher. This thesis uses modelled pollution concentration data, which were obtained from the Department for Environment, Food and Rural Affairs (DEFRA). Modelled estimates for each 1 km square grid across

the United Kingdom are produced, then to get the aggregate measurement at the intermediate geography level the median concentration over the grid squares that lie within each intermediate geography is used. An intermediate geography is a small area, known as an administrative unit, which contains a population of around 4000 people on average. Typically a Bayesian modelling approach is used to estimate the disease risk, using the available covariate information and a set of random effects to model the spatial autocorrelation.

Therefore the aim of this thesis is to examine if allowing for residual spatial autocorrelation improves the quality of fixed effect estimation compared to ignoring the spatial autocorrelation, and what modelling approach is the best to allow for the spatial autocorrelation. These aims are examined using two simulation studies, before the modelling approaches are illustrated using a motivating example on air pollution and hospital admissions due to respiratory illness.

The rest of this thesis is split into five chapters. Chapter 2 presents a summary of the methods used within this thesis, and an examination of the existing literature on air pollution and health studies. An introduction to Bayesian methods is included, which covers Markov Chain Monte Carlo simulation. The four models used in this thesis are described, the Quasi-Poisson generalised linear model and the three different spatial autocorrelation models considered. The latter include a Conditional Autoregressive model proposed by Leroux et al. (1999), the Sparse Spatial Generalised Linear Mixed model proposed by Hughes and Haran (2013) and the Localised Conditional Autoregressive model proposed by Lee et al. (2014). The existing literature critiqued in this chapter focuses on areal unit studies of air pollution and health, such as Elliott et al. (2007), Haining et al. (2007) and Lee et al. (2009).

Chapter 3 presents a simulation study which aims to determine whether the Quasi-Poisson model is appropriate for use with spatial data. The study uses a square grid as the spatial area, and assesses the performance of a Quasi-Poisson log linear model under a range of autocorrelation scenarios for the covariate and the residuals. There are six scenarios for the autocorrelation considered, where three focus on varying autocorrelation in the residual structure with a fixed level of autocorrelation in the covariate, and the remaining three focus on varying autocorrelation in the covariate with a fixed level of autocorrelation in the residual structure. The study looks at whether the Quasi-Poisson model is appropriate in any of these autocorrelation scenarios, where the model performance is assessed through root mean square error and coverage probabilities of the fixed effect estimates and uncertainty intervals.

Chapter 4 presents a further simulation study which assesses the performance of the three spatial autocorrelation models described in Chapter 2 in comparison to a Quasi-Poisson generalised linear model. Three levels of autocorrelation are considered for the covariate and the residuals. The values considered correspond to independence, moderate and strong autocorrelation. The models are also assessed under different disease prevalences, and two values for the scale of the unmeasured confounding are also considered.

Chapter 5 analyses the motivating air pollution and health data set using the four methods compared in the previous chapter. The data set focuses on respiratory illness hospital admissions in the central belt of Scotland in 2010. The Scottish Neighbourhood Statistics database provide the respiratory health hospital admissions data along with data on covariates, and the Department for Environment, Food and Rural Affairs (DEFRA) provide the data on the pollutant levels. The analysis focuses on determining if there is any significant risk associated with increasing pollution levels on the risk

of hospital admission due to respiratory illness whilst accounting for other covariates. The pollutants of interest are Nitrogen Dioxide (NO_2, μgm^{-3}), Sulphur Dioxide (SO_2, μgm^{-3}), and PM_{10} (μgm^{-3}) and $PM_{2.5}$ (μgm^{-3}) which are measures of particulate matter in the air less than 10 and 2.5 micrometres in diameter respectively. Chapter 6 presents a final discussion of the results of the thesis, the limitations of this work and areas where there is potential for future work.

Chapter 2

Methods and Existing Literature.

2.1 Data

The data are observations from N non-overlapping areal units, $A_1, \ldots, A_n$, such as electoral wards or intermediate geographies. The response, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, are population level counts of disease cases or deaths from each areal unit, and the expected counts, $\boldsymbol{E} = (E_1, \ldots, E_n)$, are available to adjust for varying population sizes and demographies across the N areal units. The expected counts are calculated by multiplying age and sex specific incidence rates from a reference population by the corresponding age and sex specific population sizes for each area, then summing over the different age and sex groups. A vector of covariates, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is available from each areal unit, and could include variables to account for concentrations of pollutants and socio-ecoonomic deprivation, for example. Disease risk in areal unit k is denoted by R_k, and a simple estimate is the Standardised Incidence Ratio (SIR) which is defined as $R_k = SIR = \frac{Y_k}{E_k}$. This is the ratio of observed to expected counts of diseases cases or deaths. A value of R_k greater than one means that areal unit A_k has an above average risk of disease, and $R_k = 1.2$ means a 20% increased risk of disease. Conversely, if R_k is less than one the

risk of disease is less than the average, and $\hat{R}_k = 0.9$ corresponds to a 10% reduction.

2.2 Poisson Generalised Linear Model and Quasi-Likelihood

The simplest regression model for count data is a generalised linear model (GLM). In a generalised linear model each Y_k is assumed to be an independent observation from an exponential family distribution $f(.)$. The model is given by

$$\begin{aligned} Y_k &\sim f(y_k|\mu_k) \quad \text{for } k = 1, \ldots, n, \\ g(\mu_k) &= \boldsymbol{x}_k^T \boldsymbol{\beta}, \end{aligned} \tag{2.1}$$

where μ_k denotes the expected value of y_k and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ is a vector of unknown regression parameters. The linear combination of all covariates is called the linear predictor, and is related to the expected value by a known monotonic invertible link function g. For the data considered here a Poisson model is used as the data are counts, and including the expected number of disease cases as an offset term the above model simplifies to

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n \\ \ln(R_k) &= \boldsymbol{x}_k \boldsymbol{\beta} \end{aligned}$$

A Poisson distribution has a mean and variance which are equal, meaning that $\mathbb{E}[Y_k] = \text{Var}[Y_k]$. However, this may not be the case and it is likely that $\text{Var}[Y_k] > \mathbb{E}[Y_k]$, which is known as overdispersion. In this case, instead of assuming Y_k is Poisson distributed, we relax this constraint and specify a quasi-likelihood model in terms of the mean and variance by:

$$\mathbb{E}[Y_k] = E_k R_k = E_k \exp(\boldsymbol{x}_k^T \boldsymbol{\beta}), \tag{2.2}$$

$$\text{Var}[Y_k] = \alpha E_k R_k = \alpha E_k \exp(\boldsymbol{x}_k^T \boldsymbol{\beta}), \tag{2.3}$$

where α is the overdispersion parameter. If $\alpha = 1$, then a Poisson model is appropriate, if $\alpha > 1$ there is overdispersion and if $\alpha < 1$ there is underdispersion. An estimate for α is given by:

$$\hat{\alpha} = \frac{1}{n-p} \sum_{k=1}^{n} \frac{(Y_k - E_k \hat{R}_k)^2}{E_k \hat{R}_k}. \tag{2.4}$$

The overdispersion parameter, α, adjusts the variance for the estimates produced from the models considered and therefore the uncertainty intervals calculated will be adjusted to account for overdispersion.

2.3 Moran's I

Moran's I (Moran (1950)) is a statistical diagnostic which measures the strength of the spatial association among data relating to n areal units. The formula for Moran's I statistic is

$$I = \frac{n \sum_i \sum_j w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}. \tag{2.5}$$

In equation (2.5), $Y_1, \ldots, Y_n$ are measurements associated with the $k = 1, \ldots, n$ areal units and W is a binary $n \times n$ neighbourhood matrix, where $w_{ij} = 1$ if areas i and j share a common border and $w_{ij} = 0$ otherwise. A Monte Carlo permutation test can be conducted to perform a test of the hypothesis

H_0 : No spatial association in the data

H_1 : Some spatial association exists in the data.

The permutation test is performed by first calculating the observed value of I for the data. Then, the data are randomly assigned to the n areal units and the Moran's I statistic is calculated. This process is repeated say 10000 times, and the p-value is computed. This is achieved by comparing the observed value of the Moran's I statistic to the remaining 9999 to determine whether it is extreme at a 5% significance level.

2.4 Bayesian Methods

The aim of a Bayesian analysis is to learn about a parameter vector $\boldsymbol{\theta}$ using the data $\boldsymbol{y}$, which is achieved by determining its posterior distribution conditional on the observed data $\boldsymbol{y}$. The posterior distribution is given by

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{y})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{2.6}$$

which is obtained from an application of Bayes Theorem. The marginal distribution of the data, $f(\boldsymbol{y})$, is calculated as $f(\boldsymbol{y}) = \sum_\theta f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$ if $\boldsymbol{\theta}$ is discrete and $f(\boldsymbol{y}) = \int_\theta f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ if $\boldsymbol{\theta}$ is continuous. Bayesian analysis is typically based on the unnormalised posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$ which is the product of the prior distribution and the likelihood function.

The posterior distribution can be calculated using various techniques depending on the complexity of the likelihood and the prior. Direct methods are available to use, however most problems are too complex to use direct methods and so approximate iterative techniques are required. The most popular of these approximate iterative techniques is Markov Chain Monte Carlo

(MCMC) simulation, which is based on a Markov chain, $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}, \ldots\}$, whose target distribution is the joint posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$. The chain is initialised by a starting value $\boldsymbol{\theta}^{(0)}$, and is run until it has converged to its target distribution. After convergence has been reached the Markov chain is sampling from the posterior distribution and as many samples as required can be generated. Markov chain simulation can be implemented using the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)) or the Gibbs sampler (Smith and Roberts (1993)), both algorithms have a target distribution equal to the posterior distribution of interest and partition the parameter vector into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ where a single iteration sequentially updates each of the d blocks in turn. The Metropolis-Hastings algorithm is the most general of these.

Algorithm 1 Metropolis-Hastings

1. Draw a starting point for the Markov chain $\boldsymbol{\theta}^{(0)}$, ensuring that its posterior probability is positive, that is $f(\boldsymbol{\theta}^{(0)}|\boldsymbol{y}) > 0$.

2. For each iteration $j = 1, 2, \ldots$ carry out steps (a) and (b) for each of the d sub-vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$. For sub-vector $\boldsymbol{\theta}_k$:

(a) Generate a possible sample $\boldsymbol{\theta}_k^*$ from a proposal distribution $q(\boldsymbol{\theta}_k^{(j)}, \boldsymbol{\theta}_k^*)$, that is based on the current value of the chain.

(b) Accept $\boldsymbol{\theta}_k^*$ as the next iteration, that is set $\boldsymbol{\theta}_k^{(j+1)} = \boldsymbol{\theta}_k^*$, with probability

$$r = \min\left\{1, \frac{f(\boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)})}{f(\boldsymbol{\theta}^{(j)}|\boldsymbol{y})q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^*)}\right\},$$

and reject it (that is set $\boldsymbol{\theta}_k^{(j+1)} = \boldsymbol{\theta}_k^{(j)}$) with probability $1-r$. In calculating the acceptance probabiliity $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(j)}$ are identical except at sub-vector k.

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm in which the proposal distribution is given by $q(\boldsymbol{\theta}_k^{(j)}, \boldsymbol{\theta}_k^*) = f(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}^{(j)}, \boldsymbol{y})$, the full conditional distribution of $\boldsymbol{\theta}_k$. For Gibbs sampling the acceptance

probability simplifies to one. Markov chain simulation is complex to implement, and the results can be affected by the choice of starting distribution, partition of the parameter vector, proposal distribution and the desired acceptance rates.

In order to compare between different Bayesian models the Deviance Information Criterion (DIC, Spiegelhalter et al. (2002)) is used. This criterion is based on a trade off between the goodness of fit of the model to the data and the complexity of the model. The deviance of the model is defined as $D(\boldsymbol{\theta}) = -2\ln[p(\boldsymbol{y}|\boldsymbol{\theta})]$, which measures the goodness of fit of the model. The complexity of the model is measured by the effective number of parameters, $p_D = \mathbb{E}[D(\boldsymbol{\theta}|\boldsymbol{y})] - D(\mathbb{E}[\hat{\boldsymbol{\theta}}|\boldsymbol{y}]) = \mathbb{E}[D(\boldsymbol{\theta}|\boldsymbol{y})] - D(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}]$. The DIC is then defined as

$$\begin{aligned} DIC &= D(\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}]) + 2p_D \\ &= \mathbb{E}[D(\boldsymbol{\theta})|\boldsymbol{y}] + p_D. \end{aligned} \tag{2.7}$$

DIC is seen as a flexible model fit statistic as it can be compared across different models and methods as long as the dependent variable is the same. Models with a smaller DIC are preferred.

2.5 Conditional Autoregressive (CAR) Models

The quasi-likelihood or Poisson models described above assume the observations are independent, which is unlikely to be realistic. Therefore one can extend the model to

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k) \\ \ln(R_k) &= \boldsymbol{x}_k^T \boldsymbol{\beta} + \boldsymbol{\phi}_k. \end{aligned} \tag{2.8}$$

Here $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are a vector of random effects, which allow for overdispersion and spatial autocorrelation in the data. They can be modelled by a conditional autoregressive (CAR) model, which is a type of Gaussian Markov random field (Besag (1974), Rue and Held (2005)). There are different types of CAR prior which include the intrinsic model, the convolution model and the Leroux model. The intrinsic model, proposed by Besag et al. (1991), is the simplest and forms the basis for the other CAR priors. The intrinsic model is represented by a multivariate Gaussian distribution:

$$\boldsymbol{\phi} \sim \text{N}(\mathbf{0}, \tau^2[\text{diag}(W\mathbf{1}) - W]^{-1}), \tag{2.9}$$

where W is the binary $n \times n$ neighbourhood matrix and $\mathbf{1}$ is a vector of ones. Its full conditional distributions $f(\phi_k|\boldsymbol{\phi}_{-k})$ are given by:

$$\phi_k|\boldsymbol{\phi}_{-k}, W, \tau^2 \sim \text{N}\left(\frac{\sum_{j=1}^n w_{kj}\phi_j}{\sum_{j=1}^n w_{kj}}, \frac{\tau^2}{\sum_{j=1}^n w_{kj}}\right), \tag{2.10}$$

where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$. The intrinsic model can only model strong correlation and can enforce too much spatial smoothness on the random effects. The joint distribution is also improper as the precision matrix, $Q = \text{diag}(W\mathbf{1}) - W$, is singular as its row sums equal zero.

One extension to this model was proposed by Leroux et al. (1999), and is represented by the multivariate Gaussian distribution

$$\boldsymbol{\phi}|W, \tau^2, \rho \sim \text{N}(\mathbf{0}, \tau^2[\rho(\text{diag}(W\mathbf{1}) - W) + (1-\rho)I_n]^{-1}). \tag{2.11}$$

The prior has a constant mean of zero and precision matrix $Q_L = \rho(\text{diag}(W\mathbf{1}) - W) + (1-\rho)I_n$ where I_n is an $n \times n$ identity matrix. The precision matrix

is a weighted average of spatially dependent (represented by $\text{diag}(W\mathbf{1}) - W$) and independent (represented by I_n) correlation structures, where the weight is equal to ρ. The univariate full conditional distributions are given by:

$$\phi_k | \boldsymbol{\phi}_{-k}, W, \tau^2, \rho \sim \text{N}\left(\frac{\rho \sum_{j=1}^{n} w_{kj}\phi_j}{\rho \sum_{j=1}^{n} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{n} w_{kj} + 1 - \rho} \right). \quad (2.12)$$

The Leroux model can represent a range of weak and strong spatial autocorrelation structures, where $\rho = 0$ corresponds to independence in space with mean 0 and constant variance. The joint distribution is proper if $0 \leqslant \rho < 1$, while $\rho = 1$ corresponds to the improper intrinsic model.

Inference for the extended model (2.8) with a set of random effects modelled by a conditional autoregressive model is typically implemented in a Bayesian setting, using Markov Chain Monte Carlo simulation. In this thesis the estimation is done in this framework, using the CARBayes (Lee (2013)) package from the statistical software R (R Core Team (2013)).

2.6 Sparse Spatial Generalised Linear Mixed Model (SGLMM)

The Sparse SGLMM was proposed by Hughes and Haran (2013) with the aim to overcome the shortcomings with existing models. Specifically, it aims to overcome problems with variance inflation due to potential spatial confounding between the random effects and any spatially smooth covariate included in $\boldsymbol{X}$, and the computational challenges posed by fitting high-dimensional latent variables in the model.

Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ and $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{H}$, where $\boldsymbol{P}$ is the residual projection matrix from a normal linear model with covariate matrix $\boldsymbol{X}$, then $\boldsymbol{P}$ and $\boldsymbol{X}$ are orthogonal. An eigendecomposition of $\boldsymbol{PWP}$ is performed

and a matrix $\boldsymbol{M}$ is created. The matrix $\boldsymbol{M}$ contains q columns of eigenvectors of $\boldsymbol{PWP}$ which correspond to the q largest positive eigenvalues. The eigenvectors of $\boldsymbol{PWP}$ correspond to all possible mutually distinct patterns of clustering residual to the covariates $\tilde{\mathbf{Z}}$ whilst accounting for the spatial structure in the data via $\mathbf{W}$. Furthermore, the eigenvectors for all positive eigenvalues correspond to positive spatial correlation that is orthogonal to $\tilde{\mathbf{Z}}$. In this way negative spatial dependence is not allowed for, as negative correlation patterns are contained in the eigenvectors of $\boldsymbol{PWP}$ that correspond to the negative eigenvalues. The magnitude of the jth eigenvalue λ_j also determines the relative importance of the spatial pattern in the jth eigenvector. The equation for the first stage of the model is:

$$g\{\mathbb{E}[Z_k|\boldsymbol{\beta}, \boldsymbol{\delta_S}]\} = \boldsymbol{x}_k^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{m}_k^\mathsf{T}\boldsymbol{\delta_S}, \tag{2.13}$$

where $\boldsymbol{\phi}_k$ in (2.8) has been replaced by $\boldsymbol{m}_k^\mathsf{T}\boldsymbol{\delta_S}$ and the prior for the random effects $\boldsymbol{\delta_S}$ becomes

$$\boldsymbol{\delta_S} \sim \mathrm{N}(0, \tau^2\boldsymbol{Q}_S^{-1}) \tag{2.14}$$

where $\boldsymbol{Q_S} = \boldsymbol{M^\mathsf{T}QM}$ is the precision matrix. To implement this method the ngspatial (Hughes and Cui (2013)) package for the statistical software R (R Core Team (2013)) is used.

2.7 Localized Conditional Autoregressive (LCAR) Model

The LCAR model was proposed by Lee et al. (2014) and seeks to improve the estimation performance of the covariate effects compared to using the traditional CAR models, as described previously. This improvement is due to a more flexible spatial autocorrelation model for $\boldsymbol{\phi}$, which can capture more realistic spatial structures that are likely to be observed in real

data. Specifically, this model is flexible spatially as it can model areas of spatial smoothness and is able to capture step changes in the random effects surface. This is achieved by allowing the random effects in the adjacent areas to be either autocorrelated or conditionally independent. The partial autocorrelation between (ϕ_i, ϕ_j) from the intrinsic CAR model is given by

$$\text{Corr}[\phi_i, \phi_j | \boldsymbol{\phi}_{-ij}, \mathbf{W}] \;=\; \frac{w_{ij}}{\sqrt{(\sum_{l=1}^{n} w_{il})(\sum_{l=1}^{n} w_{jl})}}, \tag{2.15}$$

which shows that areas that are geographically adjacent (denoted $i \sim j$) with $w_{ij} = 1$ are correlated. In contrast, for non-adjacent areas where $w_{ij} = 0$ the random effects are conditionally independent. The proposal here is to allow $\mathcal{W} = \{w_{ij} | i \sim j\}$ to be estimated from the data as binary random quantities, thus allowing neighbouring random effects to be conditionally independent or correlated. However, (2.10) shows that in this case if an area has all its w_{ij} elements estimated as zero then its conditional mean and variance are infinite, which is inappropriate. Therefore, Lee et al. (2014) use an extended vector of random effects $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \phi_*)$, where ϕ_* is a global random effect preventing the infinite conditional mean and variance problem. The $(n+1) \times (n+1)$ neighbourhood matrix for $\tilde{\boldsymbol{\phi}}$ is

$$\tilde{\mathbf{W}} \;=\; \begin{bmatrix} \mathbf{W} & \mathbf{w}_* \\ \mathbf{w}_*^{\mathrm{T}} & 0 \end{bmatrix}, \tag{2.16}$$

where $\mathbf{w}_* = (w_{1*}, \ldots, w_{n*})$ and $w_{k*} = \text{I}[\sum_{i \sim k}(1 - w_{ki}) > 0]$. Here I[.] denotes an indicator function, so that $w_{k*} = 1$ if at least one element w_{kj} relating to areal unit $\mathcal{A}_k$ has been estimated as zero, otherwise w_{k*} equals zero. Based on this an intrinsic CAR prior is specified for $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \phi_*)$, which has conditional mean and variance given by

$$\begin{aligned} \phi_k | \tilde{\boldsymbol{\phi}}_{-k} \;&\sim\; \text{N}\left(\frac{\sum_{i=1}^{n} w_{ki}\phi_i + w_{k*}\phi_*}{\sum_{i=1}^{n} w_{ki} + w_{k*} + \epsilon}, \frac{\tau^2}{\sum_{i=1}^{n} w_{ki} + w_{k*} + \epsilon} \right) \\ &\quad \text{for } k = 1, \ldots, n, \\ \phi_* | \tilde{\boldsymbol{\phi}}_{-*} \;&\sim\; \text{N}\left(\frac{\sum_{i=1}^{n} w_{i*}\phi_i}{\sum_{i=1}^{n} w_{i*} + \epsilon}, \frac{\tau^2}{\sum_{i=1}^{n} w_{i*} + \epsilon} \right). \end{aligned} \tag{2.17}$$

Here ϵ is a constant included to allow the precision matrix to be diagonally dominant and hence invertible, as the determinant has to be computed in the MCMC updating scheme. However, estimating the set $\mathcal{W}$ as binary random quantities is problematic, due to the large number of parameters. The dimensionality of $\mathcal{W}$ is $N_{\mathcal{W}} = \mathbf{1}^T \boldsymbol{W} \mathbf{1}/2$, and due to the binary nature of each edge the sample space has a size of $2^{N_{\mathcal{W}}}$. Therefore $\mathcal{W}$ is considered as a single random quality with the following prior for its neighbourhood matrix representation $\tilde{\boldsymbol{W}}$:

$$\tilde{\boldsymbol{W}} \sim \text{discrete uniform}(\tilde{\boldsymbol{W}}^{(0)}, \tilde{\boldsymbol{W}}^{(1)}, \ldots, \tilde{\boldsymbol{W}}^{(N_{\mathcal{W}})}). \tag{2.18}$$

The last candidate value $\tilde{\boldsymbol{W}}^{(N_{\mathcal{W}})}$ sets all $N_{\mathcal{W}}$ elements $w_{ij} = 1$ and corresponds to the intrinsic autoregressive model, proposed by Besag et al. (1991) described previously, for global spatial smoothing. Moving from $\tilde{\boldsymbol{W}}^{(j)}$ to $\tilde{\boldsymbol{W}}^{(j-1)}$ sets one additional $w_{kj} = w_{jk} = 0$. Therefore $\tilde{\boldsymbol{W}}^{(0)}$ contains only zeros and corresponds to independent random effects. The set $\{\tilde{\boldsymbol{W}}^{(j)} | j = 1, \ldots, N_{\mathcal{W}} - 1\}$ corresponds to localized spatial smoothing where there are some elements $w_{ij} = 1$ in the model and the corresponding random effects are smoothed, while other elements $w_{ij} = 0$ and there is no such smoothing. The set of $N_{\mathcal{W}}$ elements in the discrete uniform prior are estimated from disease data for preceding years, as it should have a similar spatial pattern in disease risk compared with the study data. The elicitation of the $N_{\mathcal{W}}$ elements in (2.18) are based on a Gaussian approximation, and further details are given by Lee et al. (2014). The full LCAR model is

specified by:

$$\begin{aligned} Y_k|E_k, R_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \boldsymbol{x}_k^T \boldsymbol{\beta} + \phi_k, \\ \tilde{\boldsymbol{\phi}} &\sim \text{N}(\mathbf{0}, \tau^2 \boldsymbol{Q}(\tilde{\boldsymbol{W}}, \epsilon = 0.001)^{-1}), \\ \tilde{\boldsymbol{W}} &\sim \text{discrete uniform}(\tilde{\boldsymbol{W}}^{(0)}, \tilde{\boldsymbol{W}}^{(1)}, \ldots, \tilde{\boldsymbol{W}}^{(N_{\mathcal{W}})}), \\ \beta_j &\sim \text{N}(0, 1000) \quad \text{for } j = 1, \ldots, p, \\ \tau^2 &\sim \text{uniform}(0, 1000), \end{aligned} \tag{2.19}$$

and is fitted using MCMC simulation.

2.8 Model Performance

To summarise a models estimation performance for covariate effects two measures are used in this thesis. These measures are the root mean square error (RMSE) of the $\hat{\beta}$ estimate of the true parameter β, and the coverage probability of its 95% uncertainty interval. For a given model the coverage is defined as the probability that the true value β is contained in the 95% uncertainty interval.

RMSE is a measure of the difference between the value of an estimator and the actual value of this parameter. The RMSE for $(\beta, \hat{\beta})$ is defined as:

$$\text{RMSE}(\hat{\beta}) = \sqrt{\mathbb{E}[(\hat{\beta} - \beta)^2]}. \tag{2.20}$$

This can be estimated by simulation, using say 500 data sets. This gives 500 estimated values of $\{\hat{\beta}^{(i)}\}_{i=1}^{500}$ from the 500 simulated data sets, and the RMSE is:

$$\hat{\text{RMSE}}(\hat{\beta}) = \sqrt{\frac{1}{500} \sum_{i=1}^{500} (\hat{\beta}^{(i)} - \beta)^2}. \tag{2.21}$$

Here, smaller RMSE indicates a model with better estimation properties. 95% uncertainty intervals for $\hat{\beta}$ are calculated differently depending on the model, and for the CAR and Sparse SGLM models 95% credible intervals are constructed by taking the 2.5% and 97.5% points of the MCMC posterior samples of β. For the quasi-likelihood model 95% confidence intervals are calculated as

$$\hat{\beta} \pm t_{n-1,0.975} se(\hat{\beta}), \tag{2.22}$$

where se($\hat{\beta}$) is its standard error. As n is large in this study, $t_{n-1,0.975}$ is 1.965927. The percentage of times that the true value of β lies in these uncertainty intervals is calculated. Due to the definition of a 95% uncertainty interval, this should be close to 95%.

2.9 Existing Areal Unit Studies on Air Pollution and Health

There are three main study designs used in air pollution and health studies, time series, cohort and areal unit, the latter of which is the focus of this thesis. Pope (1991) and Ruidavets et al. (2005) give examples of time series designs which are summarised here, and the study by Nafstad et al. (2004) is a cohort study that is summarised. However, the main part of this literature review focuses on areal unit studies.

Pope (1991) presents a time series study looking at the effect of short term exposure on health. This study assesses the association between respiratory hospital admissions and PM_{10} pollution in Utah, Salt Lake and Cache Valleys in the period between April 1985 and March 1989. In Utah Valley there was an integrated steel mill which was in operation for part of the

study period and accounted for high proportions of the valley's industrial PM_{10} pollution and total PM_{10} pollution. Comparisons were made for Utah, Salt Lake and Cache Valleys to determine if the respiratory admissions were most affected in Utah Valley. Monthly hospital admissions were regressed on a trend variable, mean monthly PM_{10} level, lagged mean monthly PM_{10} level, monthly mean low temperature and lagged monthly mean low temperature. Autoregressive regression models were estimated using maximum-likelihood estimation. Pope (1991) found that hospital admissions in Utah Valley were higher when the steel mill was open. The results of the regression analysis showed statistically significant correlations between respiratory hospital admissions and monthly mean PM_{10} levels in Utah and Salt Lake Valleys.

Ruidavets et al. (2005) presents a time series study which considers the concentrations of three pollutants, SO_2, NO_2 and O_3, and the association between these pollutants and the occurence of acute myocardial infarction from January 1997 to June 1999 in the southwest of France. A conditional logistic regression analysis was performed to calculate the relative risk. The relative risk for an increase of O_3 concentration for occurence of acute myocardial infarction were significant for current day and one day lag measurements. NO_2 and SO_2 exposures were not significantly asssociated with acute myocardial infarction.

Nafstad et al. (2004) presents a cohort study to look at the effects of long term air pollution exposure considers men ages 40 to 49 living in Oslo, Norway in 1972. These men were followed from 1972-1973 until 1998. Indicators of air pollution were sulphur dioxide and nitrogen oxides and the health outcomes considered were total deaths from diseases, deaths from respiratory diseases, deaths from lung cancer, deaths from ischemic heart diseases and deaths from cerebrovascullar diseases. A Cox proportional hazard regression model was used to evaluate the association between deaths and the indica-

tors of air pollution. The results showed that NO_x exposure was associated with the risk of dying in this cohort of men.

Jerrett et al. (2005) focuses on total suspended particulate data from 23 monitoring stations which are interpolated using kriging. The analysis considers the study region of census tracts of Hamilton, Canada, and mortality which considered all causes minus traumatic deaths, premature mortality, deaths due to cardio-respiratory problems and deaths due to all causes of cancer. The results found show that long term exposure to particulate air pollution is associated with all cause, premature, cardio-respiratory and cancer mortality for males in Hamilton. For females in Hamilton, there is an association with all cause, premature and cancer mortality.

Maheswaran et al. (2005) look at 1030 census enumeration districts in Sheffield as the unit of analysis and consider stroke deaths and hospital admissions from 1994 to 1998. The pollutants which the paper uses are PM_{10}, NO_x and CO. Poisson regression methods were used and spatial autocorrelation in model residuals between neighbouring areas are accounted for using a conditional autoregressive spatial model. The results found that increasing outdoor air pollution levels were significantly associated with increasing stroke mortality risk at the small area levels.

Dominici et al. (2006) looks at the risk of cardiovascular and respiratory hospital admissions assocciated with $PM_{2.5}$. The analysis in this paper is based on daily counts of hospital admissions for 1999 to 2002. A bayesian two stage hierarchical model is used. The results found show a short term increase in hospital admission rates associated with this pollutant for all cardiovascular and respiratory health outcomes considered.

Elliott et al. (2007) look at associations between black smoke and SO_2 and mortality across electoral wards in Great Britain. The exposure indices were averaged over four year periods from April 1996 - March 1970 to April 1990 - March 1994 to smooth year to year fluctuations in concentrations and minimise the effects of missing data. Mortality data was summed over four successive 4 year mortality periods from April 1982 - March 1986 to April 1994 - March 1998. Observed deaths for each electoral ward were modelled as Poisson with a log-linear function for pollutant data and a log-normal random intercept to allow for overdispersion in the Poisson model, where the independence and correlation models are based on a geostatistical model. Models were fitted within a Bayesian framework using Markov Chain Monte Carlo methods. Significant associations were found between black smoke and SO_2 concentrations and mortality. The effect was stronger for respiratory illness than other causes of death for the most recent exposure periods and the most recent mortality period.

Haining et al. (2007) uses small area data from Sheffield in the 1994 to 1998 period to investigate the effects of NO_x pollution on coronary heart disease mortality. Models were fitted using a Bayesian approach and the two types of models compared were a generalised Poisson log-linear model with adjustment for overdispersion and a hierarchical Poisson log-linear model with spatial random effects specified using a CAR model. The results show significant effects of NO_x on coronary heart disease mortality using both models, although the effects are not significant at all NO_x categories depending on what the model also adjusts for and which of the models is used.

Lee et al. (2009) considers four major urban areas, Aberdeen, Dundee, Edinburgh and Glasgow, and the associations between respiratory hospital admissions in 2005 and exposure to PM_{10} and NO_2. Inference is implemented within a Bayesian framework using MCMC simulation. The results show that

PM_{10} and NO_2 are significantly associated with respiratory hospital admissions in Edinburgh and Glasgow, whereas there is not a significant increase in risk for Aberdeen and Dundee.

Lee and Mitchell (2014) compare three models for the effect of air pollution on respiratory health which are a overdispersed Poisson GLM, and two CAR models the Besag-York-Mollie (BYM) model and the localised smoothing model. In the paper, four types of air pollutant, namely PM_{10}, $PM_{2.5}$, NO_2 and CO, are considered for the Greater Glasgow and Clyde health board area between 2007 and 2011. The results for the localised smoothing model show that increases in NO_2 and PM_{10} are associated with an increased risk of respiratory ill health.

Chapter 3

Investigating The Effects of Ignoring Residual Spatial Autocorrelation on Fixed Effect Estimates.

3.1 Introduction

In a number of epidemiological studies there is spatial autocorrelation within the data, where observations from areas closer together are more similar than areas further apart. This autocorrelation often remains in the residuals after any known covariate effects have been accounted for, and is known as residual spatial autocorrelation. This residual spatial autocorrelation can be induced by a number of factors, and violates the assumption of independence that is common in many regression models. One possible cause is unmeasured confounding, which occurs when an important spatially autocorrelated covariate is either unmeasured or unknown. The spatial structure in this covariate induces spatial autocorrelation into the response, which hence cannot be accounted for in a regression model. Other possible causes of residual spatial autocorrelation are neighbourhood effects, where subjects' behaviour

is influenced by that of neighbouring subjects, and grouping effects, where subjects choose to be close to similar subjects. In order to analyse the data appropriately, which could include data on respiratory disease in Scotland and air pollution data for example, the spatial autocorrelation should be accounted for. If this autocorrelation is not taken into account then any results produced may be inappropriate. However, the magnitude of the possible poor performance of models based on the assumption of independence when applied to spatially autocorrelated data is unknown. Specifically, the quality of the estimation of these models is likely to depend on numerous factors, such as the level of residual spatial autocorrelation in the data, as well as whether the covariates included are themselves spatially autocorrelated. This chapter seeks to adddress these questions empirically, via a large simulation study. This study uses a square grid as the spatial area of interest, and assesses the performance of a Quasi-Poisson log linear model which assumes the residuals are independent. The root mean square error and coverage probabilities of the estimated covariate effects are considered to summarise the model quality. This chapter aims to answer whether Quasi-Poisson models are appropriate in a range of circumstances where there is autocorrelation in the covariate and the residuals at a variety of levels.

3.2 Data Generation

The study region is a square spatial grid of dimension 20×20, yielding $n = 400$ areas in total, from which the distance matrix, D, and neighbourhood matrix, W, are calculated. The neighbourhood matrix W is a binary $n \times n$ matrix where $w_{ij} = 1$ if areas i and j share a common border and $w_{ij} = 0$ otherwise. D is an $n \times n$ distance matrix, where d_{ij} is the Euclidean distance between areas i and j. A total of 500 simulated data sets are gener-

ated under a number of different scenarios in this study, to ensure the results are robust to random chance.

For $k = 1, \ldots, n$, data are generated under the following model:

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n, \\ \ln(R_k) &= \beta_0 + \beta_1 x_k + \phi_k, \end{aligned} \tag{3.1}$$

where Y_k is the observed number of cases of disease, E_k is the expected number of cases of disease and R_k is the risk. Throughout this chapter, the true values for β_0 and β_1 are specified as 0 and 1 respectively. We consider three different scenarios for the expected number of cases, the expected number of cases in each area is fixed at either 10, 100 or 1000. For each simulated data set the covariate, $\underline{\boldsymbol{x}} = (x_1, \ldots, x_n)$, and the random effects $\underline{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_n)$ are regenerated so that the results are not affected by the particular realisation chosen.

We consider two different scenarios for the covariate $\underline{\boldsymbol{x}} = (x_1, \ldots, x_n)$, either $\underline{\boldsymbol{x}}$ is independent in space or it is autocorrelated in space. If $\underline{\boldsymbol{x}}$ is autocorrelated in space then $\underline{\boldsymbol{x}}$ is generated as $\underline{\boldsymbol{x}} \sim \text{Normal}(\underline{\mathbf{0}}, \tau^2\Lambda)$, where τ^2 is kept fixed at 1 and $\Lambda = [\rho(\text{diag}(W1) - W) + (1-\rho)I]^{-1}$. Here, Λ is the variance matrix corresponding to the conditional autoregressive model proposed by Leroux et al. (1999) and is commonly used to model spatial autocorrelation. The parameter ρ must be in the interval [0, 1), where a value close to 1 indicates strong autocorrelation, whilst if ρ equals 0 then the covariate is uncorrelated.

The random effects $\underline{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_n)$ induce spatial autocorrelation and overdispersion into the simulated disease counts $\underline{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)$ and are generated as $\underline{\boldsymbol{\phi}} \sim \text{Normal}(\underline{\mathbf{0}}, \sigma^2\Sigma)$. Here, $\Sigma = \exp(-\nu D)$, where D is the $n \times n$ distance matrix and the range parameter $\nu \in (0, \infty)$ affects the auto-

correlation. When ν is small there is strong autocorrelation, whilst if ν is large then there is weak autocorrelation present in the random effects. The value of σ^2 is fixed at 1. In each scenario $\underline{\boldsymbol{\phi}}$ is standardised, to have mean zero and variance one. However to assess the impact of the magnitude of the random effects on model performance, multiples 1, 0.1 and 0.01 of $\underline{\boldsymbol{\phi}}$ are considered. We note that different mechanisms were used to induce spatial structure into $\underline{\boldsymbol{\phi}}$ and $\underline{\boldsymbol{x}}$, as they are likely to be generated by different underlying processes.

A Quasi-Poisson model is fitted to the simulated data. A Poisson distribution has a mean and variance which are equal, meaning that $\mathbb{E}[Y_k] = \text{Var}[Y_k]$. However, this may not be the case and it is likely that $\text{Var}[Y_k] > \mathbb{E}[Y_k]$, which is known as overdispersion. In this case, instead of assuming Y_k is Poisson distributed, we relax this constraint and specify the model in terms of the mean and variance by:

$$\mathbb{E}[Y_k] = E_k R_k, \tag{3.2}$$

$$\text{Var}[Y_k] = \alpha E_k R_k, \tag{3.3}$$

where α is the overdispersion parameter. If $\alpha = 1$, then a Poisson model is appropriate, if $\alpha > 1$ there is overdispersion and if $\alpha < 1$ there is underdispersion. An estimate for α is given by:

$$\hat{\alpha} = \frac{1}{n-p} \sum_{k=1}^{n} \frac{(Y_k - E_k \hat{R}_k)^2}{E_k \hat{R}_k}. \tag{3.4}$$

The following simulation scenarios are considered within this chapter:

- Scenario 1 - Varying autocorrelation in the residual structure with an uncorrelated covariate, which corresponds to varying ν for $\rho = 0$.

- Scenario 2 - Varying autocorrelation in the residual structure with a moderately autocorrelated covariate, which corresponds to varying ν for $\rho = 0.5$.

- Scenario 3 - Varying autocorrelation in the residual structure with a strongly autocorrelated covariate, which corresponds to varying ν for $\rho = 0.9$.

- Scenario 4 - Varying autocorrelation in the covariate with weakly autocorrelated residual structure, which corresponds to varying ρ for $\nu = 30$.

- Scenario 5 - Varying autocorrelation in the covariate with moderately autocorrelated residual structure, which corresponds to varying ρ for $\nu = 1.4$.

- Scenario 6 - Varying autocorrelation in the covariate with strongly autocorrelated residual structure, which corresponds to varying ρ for $\nu = 0.3$.

Thus the first three scenarios illustrate the performance of the Quasi-Poisson generalised linear model in the situation of a fixed covariate structure (either independent in space or autocorrelated) with varying levels of residual spatial autocorrelation. The last three scenarios consider the converse, where the level of residual spatial autocorrelation is fixed and the level of autocorrelation in the covariate is varied. A range of values of ν are considered to give a range of mean autocorrelation strengths across the study region between 0.006719 and 0.94698, while values of ρ between 0 and 0.9 are considered for the autocorrelation of the covariate in the first three scenarios.

To summarise model performance two measures are used. These measures are the root mean square error (RMSE) of the $\hat{\beta}_1$ estimate, and the coverage probability of its 95% confidence interval. The coverage is defined

as the probability that the true value β_1 is contained in the 95% confidence interval for $\hat{\beta}_1$.

RMSE is a measure of the difference between the value of an estimator and the actual value of this parameter. The RMSE is defined in Equation 2.20 and for the 500 estimated values of $\{\hat{\beta_1}^{(i)}\}_{i=1}^{500}$ from the 500 simulated data sets the RMSE is defined in Equation 2.21. A smaller RMSE indicates a better fitting model.

Confidence intervals for $\hat{\beta}_1$ are calculated as:

$$\hat{\beta}_1 \pm t_{n-1,0.975} se(\hat{\beta}_1). \tag{3.5}$$

As n is large in this study, $t_{n-1,0.975}$ is 1.965927. The percentage of times that the true value of β_1 lies in these confidence intervals is calculated. Due to the definition of a 95% confidence interval, this should be close to 95%.

3.3 Results

3.3.1 Scenario 1 - Varying correlation in the residual structure with an uncorrelated covariate

In this first scenario the covariate is uncorrelated in space, and we examine the impact on model performance (as measured by the RMSE and coverage probability) of varying levels of residual spatial autocorrelation. We note that in changing the residual spatial autocorrelation induced by the random effects the variance of the random effects will also change, as the more correlated they are the smaller their variation will be. The level of variation in the random effects will also affect model performance, as the larger it gets the

larger the size of the unmeasured confounding that is induced into the model, hence estimation performance will reduce. Therefore the random effects are standardised to have a mean of zero and a standard deviation of one, hence the only difference between the random effects for each value of ν will be their level of autocorrelation.

Figures 3.1, 3.2 and 3.3 show the RMSE (top plot) and coverage probabilities (bottom plot) for the estimated covariate effect, when the covariate is uncorrelated and the residual spatial autocorrelation is varied from independence to strong autocorrelation. The three figures differ in the scaling that is applied to the random effects, with their magnitudes in Figure 3.2 and 3.3 being divided by 10 and 100 respectively compared to their size in Figure 3.1.

Figure 3.1 shows a mean RMSE of 0.1207, and the coverage probabilities range between 61.40% and 78.40%, with a mean coverage probability of 71.30%. Adjusting the scale of the standardised ϕ by multiplying it by 0.1 produces the model quality summaries in Figure 3.2, with a mean RMSE of 0.01115 and a mean coverage of 74.68%. Finally, Figure 3.3 is produced by adjusting the scale of the standardised ϕ by multiplying it by 0.01. The RMSE in Figure 3.3 shows a mean of 0.004149 and a mean coverage of 94.29%, and the latter varies between 91.00% and 97.20%. As the size of ϕ gets bigger, the model quality summaries get worse. This is to be expected, as the magnitude of the unexplained variation contaminating the covariate effect increases. However, when the multiplier of ϕ is not very small (not 0.01), the coverage probabilities suggest that the Quasi-Poisson models cannot handle this excess variation as the coverages are well below their nominal 95% levels. Finally, the spatial autocorrelation in the residuals seems to have no effect on parameter estimation for uncorrelated covariates, as no differences are observed as ν changes.

To consider whether changing the expected number of disease cases affects the results found, two further values for the disease prevalence across the study region are examined, namely 10 or 1000 expected cases across the region and these are shown in Figures 3.4 and 3.5 respectively. The same cases of scaling of the random effects were considered as used in Figures 3.1, 3.2 and 3.3, although only the case comparable to Figure 3.1, where the magnitude is one, is shown in the Figures 3.4 and 3.5. Figures 3.4 and 3.5 show the RMSE (top plot) and coverage probabilities (bottom plot) for the estimated covariate effect, when the covariate is uncorrelated and the residual spatial autocorrelation is varied from independence to strong autocorrelation. Figure 3.4 shows a mean RMSE of 0.1215, and the coverage probabilities range between 67.20% and 76.60%, with a mean coverage probability of 71.58%. Figure 3.5 shows the RMSE has a mean of 0.1205, and a range of coverage probabilities between 63.60% and 77.40%, with a mean coverage probability of 71.30%. The results shown for each of the different number of expected cases seem similar, as they all show no pattern according to the strength of the residual spatial autocorrelation. This suggests that there is no effect due to the spatial autocorrelation in the residuals when $\underline{\boldsymbol{x}}$ is uncorrelated. The range of values for RMSE and coverage are similar in each of the different scenarios for expected number of cases considered here, so in this scenario where the covariate is uncorrelated there does not seem to be an effect from whether the disease is more or less common.

Figure 3.1: RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases, and the expected number of disease cases is equal to 100.

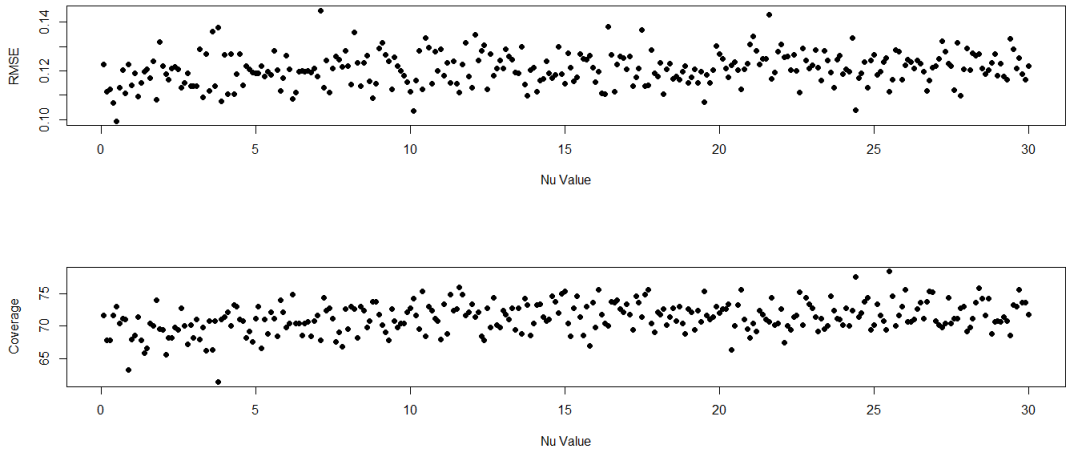


Figure 3.2: RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1.

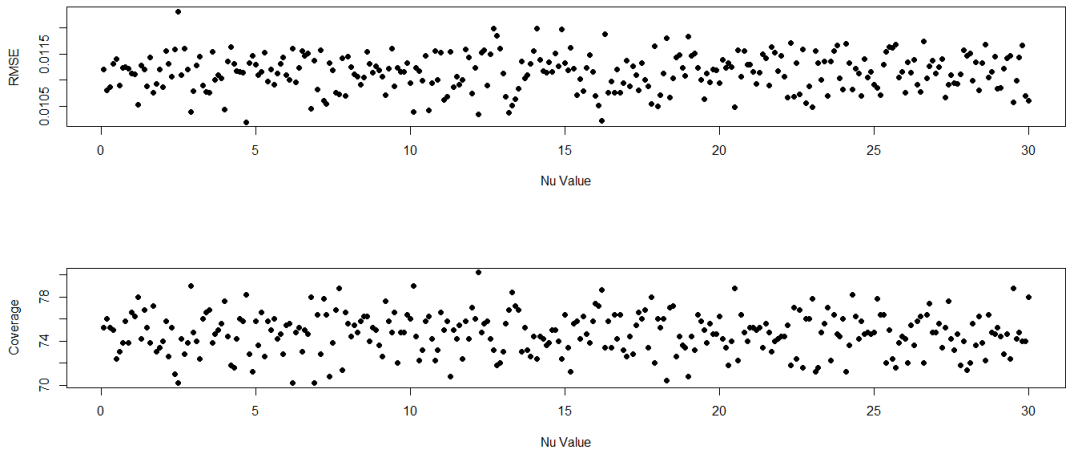


Figure 3.3: RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation varies. This correlation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01.

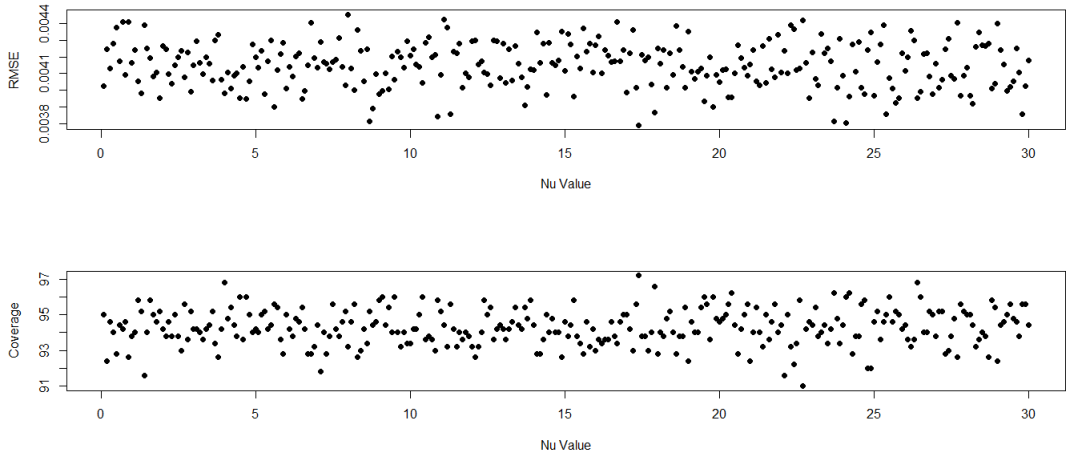


Figure 3.4: RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10.

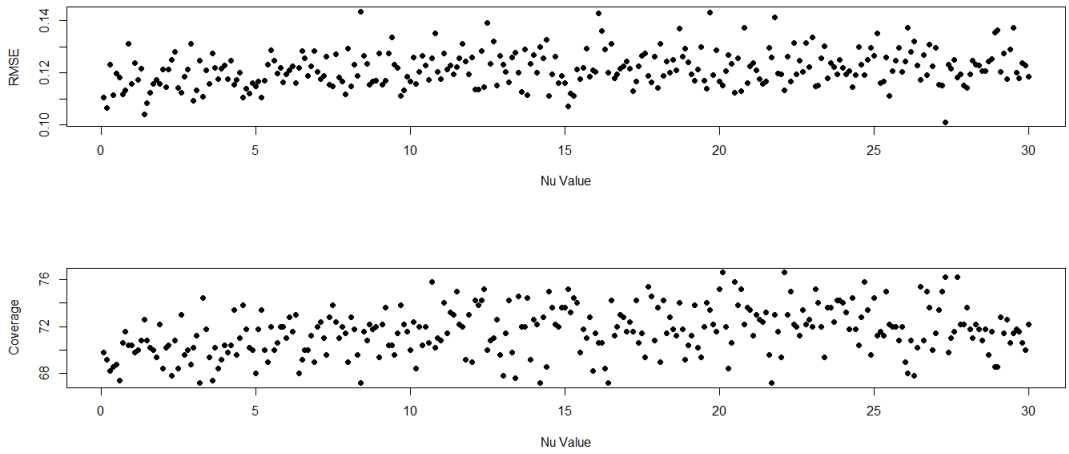
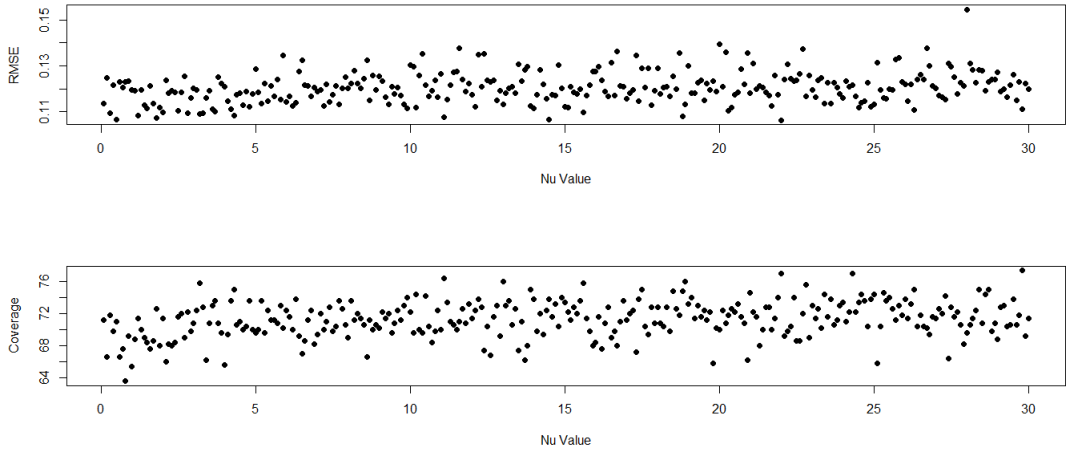


Figure 3.5: RMSE and Coverage for β_1 when the covariate is uncorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000.



3.3.2 Scenario 2 - Varying autocorrelation in the residual structure with a moderately autocorrelated covariate

In this second scenario the covariate is moderately autocorrelated in space, with $\rho = 0.5$, and we examine the impact on model performance of varying the levels of residual spatial autocorrelation. The random effects are again standardised to have a mean of zero and a variance of one. Figure 3.6 shows the RMSE decreases as the value of ν increases, meaning that the RMSE decreases as the correlation becomes weaker. The RMSE ranges from a maximum of 0.1744, when the correlation in the residual structure is highest, to a minimum of 0.1212 at the weakest correlation in the residual structure. The coverage reaches a maximum of 83.20% when the correlation in the residual structure is weakest. The minimum value of the coverage is 63.40% and this occurs when the correlation is strongest. Figure 3.7 shows the RMSE and coverage under the scenario of multiplying the standardised

ϕ by 0.1 with the case of a moderately correlated covariate. The RMSE in this scenario shows a decreasing trend as the residual correlation becomes weaker. The maximum value of RMSE is 0.01734 and the minimum value is 0.01219. The coverage shows an increasing trend, as the correlation becomes weaker, from 67.80% at a minimum to a maximum of 85.80%. Figure 3.8 shows a range of RMSE values between 0.005609 and 0.006803, and a range of coverage probabilities between 91.00% and 97.80% with a mean coverage of 94.63%. This case does not show any pattern in the effect of the correlation in the residual structure, which indicates that the scale of the random effects is too small to have any impact.

The results differ from the first scenario, and show that the presence of spatial autocorrelation in the residuals does adversely impact upon the estimation of covariate effects if that covariate itself exhibits spatial structure. For example, in Figure 3.6 the RMSE reduces by around 30% as the residual spatial autocorrelation reduces, while the coverage probability increases by around 20%. One possible reason for this is the potential for collinearity between the spatially autocorrelated covariate and the random effects, which adversely affects the estimation performance of the former. This phenomenon differs from scenario 1, and suggests that residual spatial autocorrelation in itself may not be a problem, but more its interplay with other spatially autocorrelated covariates. In common with scenario 1 however, these results suggest that unless the random effects have very small size (Figure 3.8), then the confidence intervals produced under the Quasi-Poisson model are too narrow regardless of the residual autocorrelation levels.

To consider whether changing the disease prevalence has an impact on the results found, the expected numbers of cases of 10 and 1000 are also examined in Figures 3.9 and 3.10 respectively. Figure 3.9 shows the RMSE has a decreasing trend as the autocorrelation in the residual structure becomes

weaker. The maximum value of RMSE is 0.1816, which occurs when ν is small and therefore the autocorrelation in the residual structure is strongest whilst the minimum value of the RMSE is 0.1225. The coverage ranges from a minimum of 63.80%, when the spatial autocorrelation is strongest, to a maximum of 83.80%. Figure 3.10 shows the RMSE shows a decreasing trend as the spatial autocorrelation becomes weaker. The RMSE ranges from a maximum of 0.1824 to a minimum of 0.1246. The coverage shows an increasing trend and reaches a maximum when the spatial autocorrelation is weakest. The coverage shows values between 63.20% and 82.80%.

The results show that including additional spatial autocorrelation when $\underline{\boldsymbol{x}}$ is moderately autocorrelated results in poorer estimation of covariate effects in terms of RMSE and coverage probabilities. There seems to be little effect from changing the expected numbers of disease cases as the overall pattern of RMSE and coverage probabilities are similar in the three cases of disease prevalance considered. The ranges of coverage probabilities found are similar when comparing the results from using the different values for the expected numbers of disease cases and the same magnitude of ϕ. There is an effect on parameter estimation if there is spatial autocorrelation in the residuals, in conjunction with a moderately correlated covariate, as the coverage increases and the RMSE decreases when ν increases, which means that the spatial autocorrelation gets weaker. The coverage probabilities suggest that the Quasi-Poisson model is not adequate in these circumstances as the coverages found are below the nominal 95% level, unless ϕ is very small as in Figure 3.8. As in scenario 1 there is little effect due to disease prevalence, the Quasi-Poisson model is not adequate, and increasing the magnitude of ϕ reduces the performance of the model. However, in this scenario there is an effect from the spatial autocorrelation in the residuals as there is a trend in the results as ν changes. Therefore, the residual spatial autocorrelation may not be the problem in itself but the combination of residual spatial autocor-

relation and spatially correlated covariates may be the problem.

Figure 3.6: RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases, and the expected number of disease cases is equal to 100.

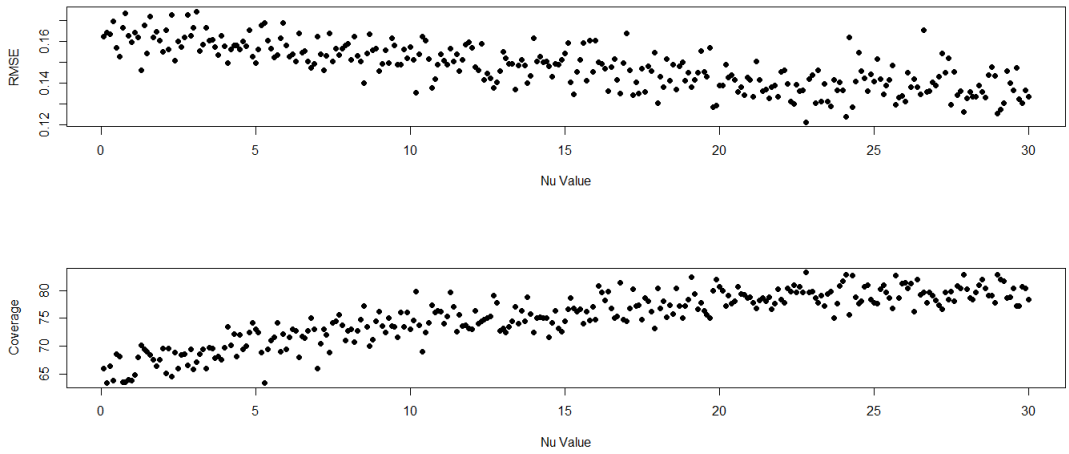


Figure 3.7: RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1.

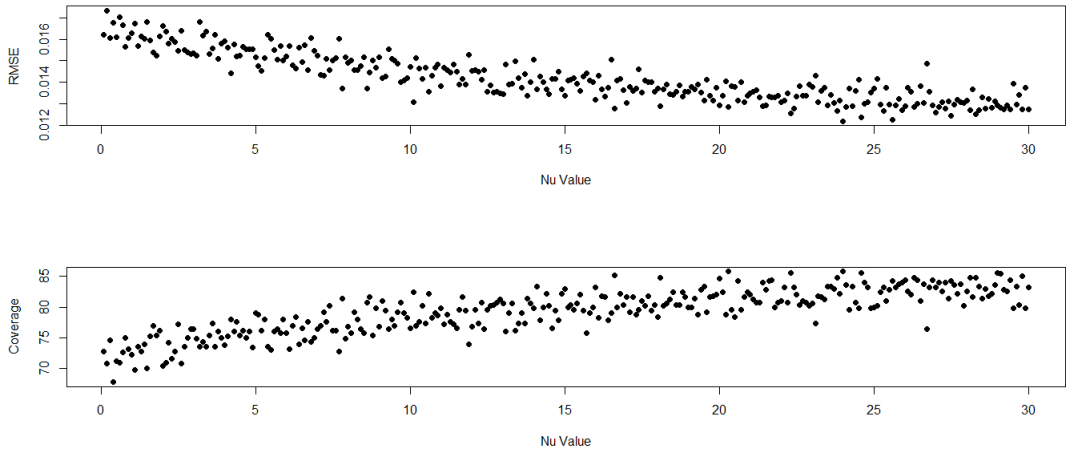


Figure 3.8: RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01.

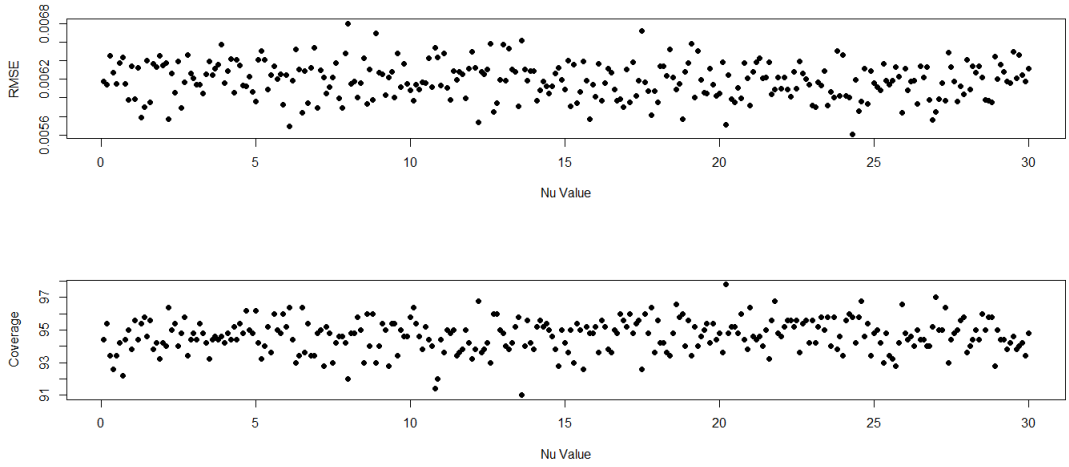


Figure 3.9: RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10.

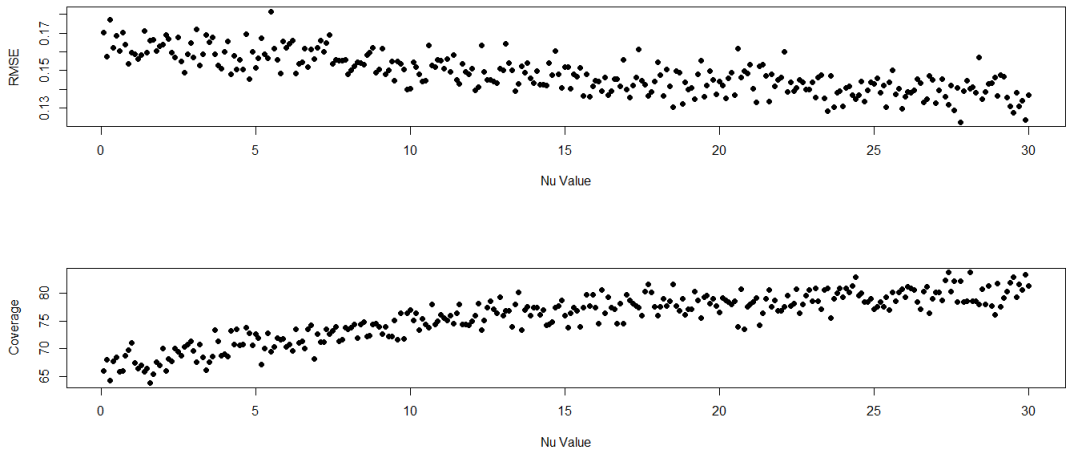
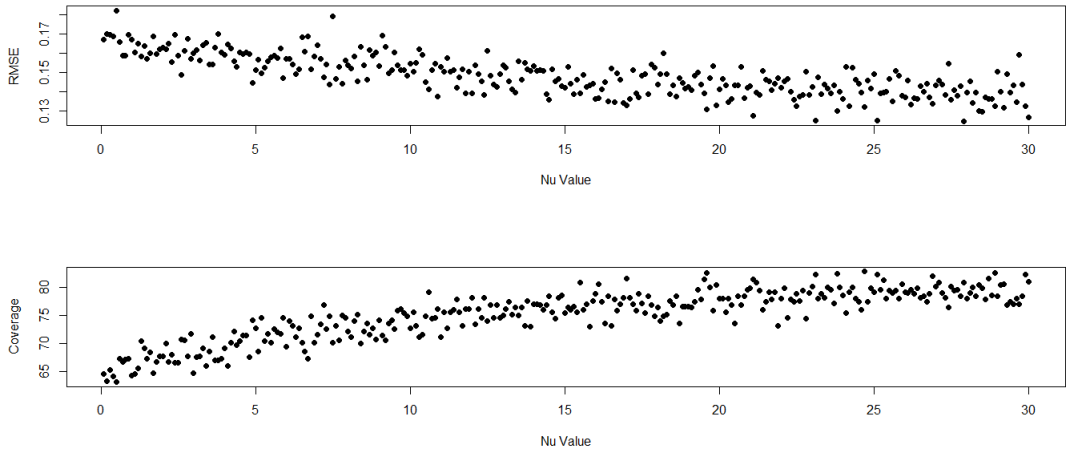


Figure 3.10: RMSE and Coverage for β_1 when the covariate is moderately autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000.



3.3.3 Scenario 3 - Varying autocorrelation in the residual structure with a strongly autocorrelated covariate

This scenario focuses on a covariate which is strongly autocorrelated in space and we examine the impact on model performance of varying the strength of the residual spatial autocorrelation. The random effects are standardised to have a mean of zero and a variance of one so the only difference between the random effects for each level of ν will be their level of autocorrelation. The Figures 3.11 to 3.15 show the RMSE (top plot) and coverage probabilities (bottom plot) for the estimated covariate effect when the covariate is strongly autocorrelated and the residual spatial autocorrelation is varied. Figures 3.11, 3.12 and 3.13 differ by the scale of the random effects as the magnitudes in Figures 3.12 and 3.13 are divided by 10 and 100 respectively compared to their size in Figure 3.11. Figure 3.11 shows the RMSE decreases as ν increases which means that the autocorrelation in the residual structure

becomes weaker. The RMSE ranges between 0.1305 and 0.2558. The coverage probabilities increase as the autocorrelation in the residual structure becomes weaker. The coverage ranges between a minimum of 40.60% and a maximum of 80.80%. Figure 3.12 shows the RMSE and coverage under the scenario of multiplying the standardised ϕ by 0.1 with a strongly autocorrelated covariate. The RMSE, shown in Figure 3.12, decreases from a maximum of 0.02543, when the autocorrelation in the residual structure is strongest, to a minimum of 0.01322, when the autocorrelation in the residual structure is weakest. The coverage shows an increasing trend from 50.00% to 83.40% as ν increases. Adjusting the magnitude of ϕ to be 0.01 of that for Figure 3.11, produces the results displayed in Figure 3.13. The RMSE ranges between 0.005698 and 0.006864, and the coverage ranges between 91.00% and 96.80%. The mean coverage in this case is 94.15%. There is only a slight trend evident in the RMSE and coverage probabilities in Figure 3.13, which indicates that the scale of the random effects is still large enough to allow them to have an impact on the results, however, it is not very strong.

In the first scenario the presence of residual spatial autocorrelation has no effect on the estimation of covariate effects, whilst the second scenario shows that residual spatial autocorrelation has an adverse impact on estimation of the covariate effects. This scenario is similar to scenario 2 as there is an impact of spatially autocorrelated residuals on the quality of estimation of covariate effects. When $\underline{\boldsymbol{x}}$ is strongly autocorrelated introducing additional spatial autocorrelation causes poorer estimation and when the magnitude of ϕ increases the results get worse. The Quasi-Poisson model seems to produce confidence intervals which are too narrow, unless the magnitude of ϕ is very small when the mean of the coverage probabilities is close to the nominal value of 95%. Therefore, all three scenarios show results which suggest that the Quasi-Poisson model is inappropriate as it cannot handle the excess variation in ϕ unless the scale of the random effects is smallest. The two

scenarios, namely scenarios 2 and 3, which show autocorrelation in the covariate indicate that spatial autocorrelation may not be a problem in itself, however, the problem may be the interplay of residual spatial autocorrelation with other spatially autocorrelated covariates. This could be due to collinearity between the spatially autocorrelated covariate with the random effects.

To consider whether changing the disease prevalence has an impact on the results found expected numbers of cases of 10 and 1000 are also examined. Considering a rare disease first, by using an expected number of disease cases of 10, produces the Figure 3.14. Figure 3.14 shows the RMSE decreases as ν increases, which means that the RMSE decreases as the residual spatial autocorrelation becomes weaker. The RMSE ranges between 0.2519, when the residual spatial autocorrelation is strongest, to 0.1332 when the residual spatial autocorrelation is weakest. The coverage shows an increasing trend as the residual spatial autocorrelation decreases. The coverage ranges from a minimum of 44.80% to a maximum of 80.60%. Changing the expected number of disease cases to 1000 produces the Figure 3.15. Figure 3.15 shows the RMSE and coverage probabilities when the expected number of disease cases is 1000. The RMSE shows as decreasing trend, whilst the coverage shows an increasing trend, as ν increases and the residual spatial autocorrelation decreases. The RMSE ranges from a maximum of 0.2552 to a minimum of 0.1305. The coverage probabilities range from 41.80% to 79.00%.

In the scenario of varying the strength of the autocorrelation with a strongly autocorrelated covariate, the disease prevalence has an impact on the results. As the expected number of disease cases increases, and so the disease becomes more common, the model performance improves. When the residual spatial autocorrelation is weakest, where ν is large, the coverage probabilities are highest and approach 95% when the magnitude of ϕ is

smallest. As the residual spatial autocorrelation decreases, ν increases, the coverage increases and the RMSE decreases. When the magnitude of ϕ is smallest, there is still an effect due to the residual spatial autocorrelation as a trend in the RMSE and coverage probabilities is still evident. The previous two scenarios found no effect from changing the disease prevalence, however, here there is some effect as shown in Figures 3.11, 3.14 and 3.15.

Figure 3.11: RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 100.

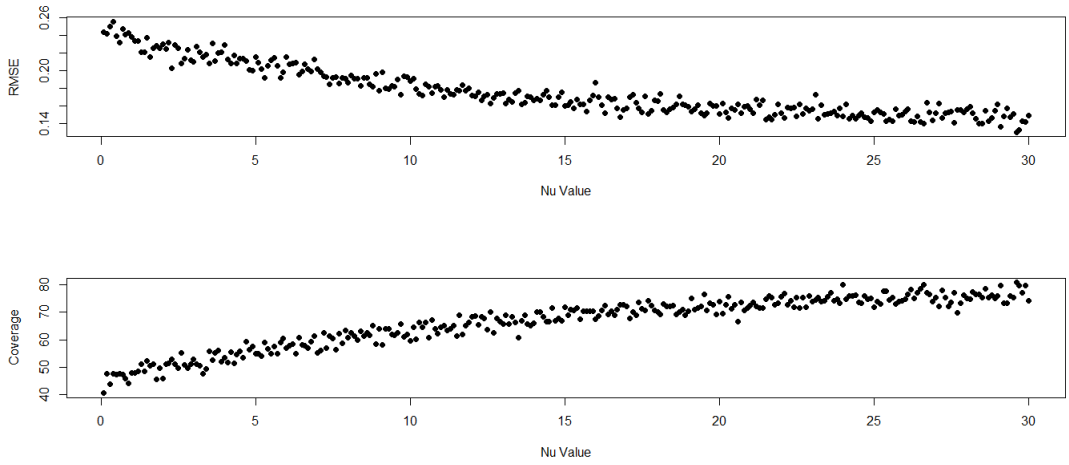


Figure 3.12: RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.1.

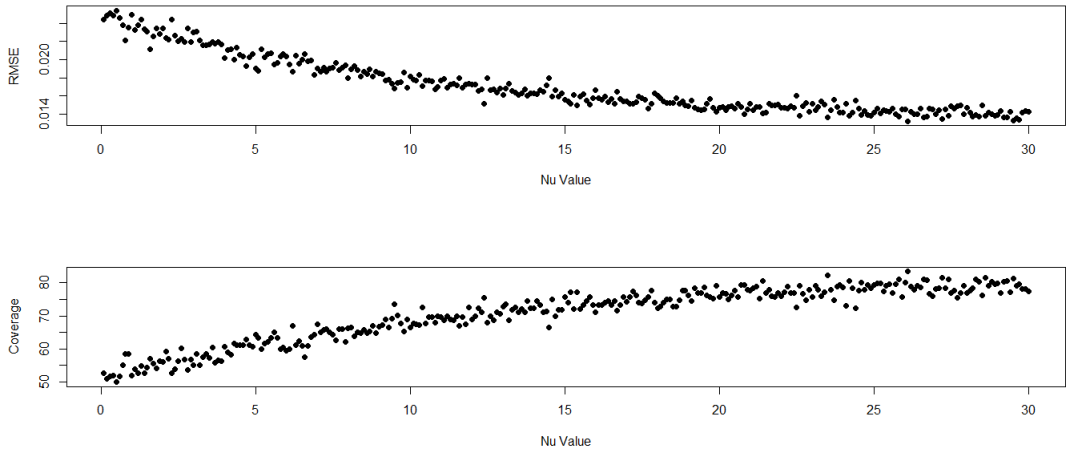


Figure 3.13: RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases with the expected number of disease cases equal to 100 and the residual spatial autocorrelation multiplied by 0.01.

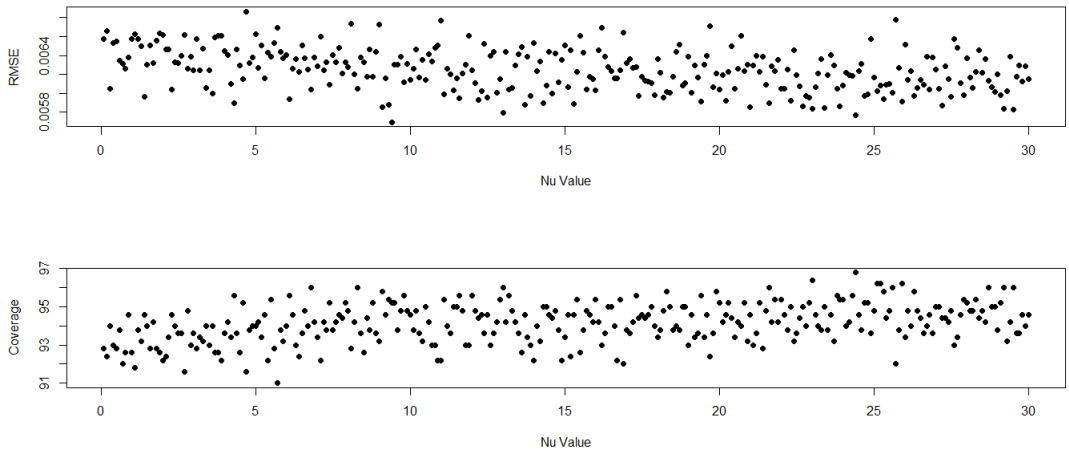


Figure 3.14: RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 10.

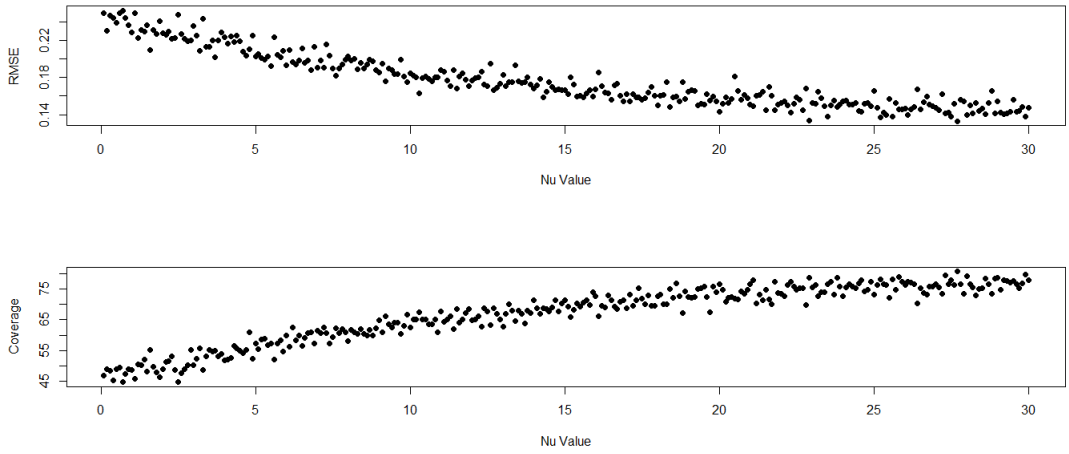
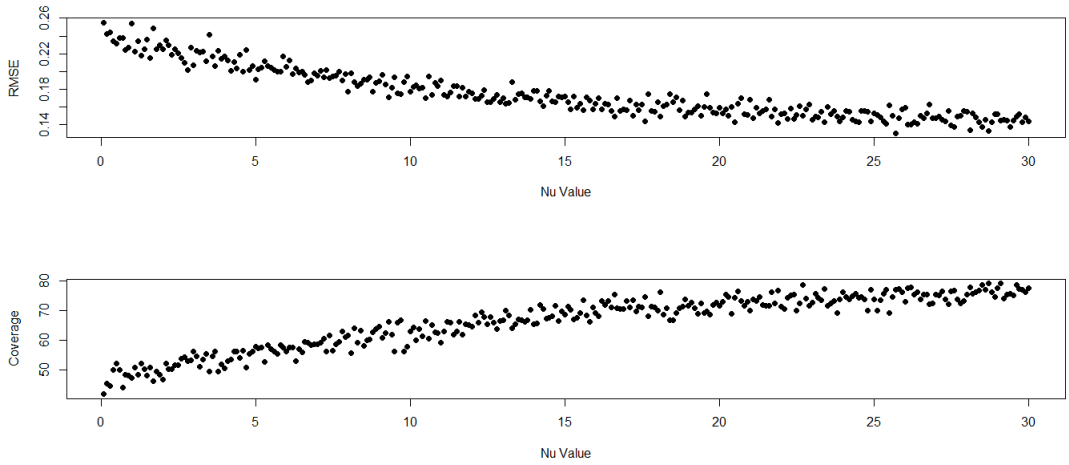


Figure 3.15: RMSE and Coverage for β_1 when the covariate is strongly autocorrelated and the residual spatial autocorrelation increases as ν decreases where the expected number of disease cases is equal to 1000.



3.3.4 Scenario 4 - Varying autocorrelation in the covariate with weakly autocorrelated residual structure

This scenario examines how changing the level of autocorrelation in the covariate, by adjusting the value of ρ, with a weakly autocorrelated residual structure impacts on model performance. The Figures 3.16, 3.17 and 3.18 show the RMSE (top plot) and coverage probabilities (bottom plot), for different expected numbers of disease cases, 100, 10 and 1000 respectively. Figure 3.16 shows a slight increase in the RMSE and a slight decrease in the coverage probabilities as ρ increases. The RMSE ranges from 0.1140 to 0.1406, with a mean value of 0.1265. The coverage probabilities range between a minimum of 64.20% to a maximum of 75.60%, with a mean coverage of 69.47%. The results for 10 expected disease cases are shown in Figure 3.17. The RMSE ranges between 0.1107 and 0.1402, with a mean of 0.1258, whilst the coverage ranges between 63.20% and 76.20%, with a mean of 70.01%. There is a slight increasing trend in the RMSE values, and a decreasing trend in the coverage probabilities, as the autocorrelation in the covariate becomes stronger. Figure 3.18 shows the RMSE and coverage when the expected number of disease cases across the region of 1000 is considered for this scenario. The RMSE ranges between 0.1116 and 0.1423, with a mean of 0.1264, where the lower RMSE values occur when the autocorrelation in the covariate is weakest. The coverage probabilities range between 63.00% and 76.40%, with a mean coverage of 69.40%.

The Figures 3.16, 3.17 and 3.18 suggest that changing the disease prevalence does not make a difference in the results in this scenario as the coverage probabilities are similar across the three cases of disease prevalence considered. The coverage probabilities are highest, around 76%, when the autocorrelation in the covariate is weakest, however this is poor as the maxi-

mum coverage is below the nominal value of 95%. Therefore even with weak autocorrelation in the residual structure the Quasi-Poisson model does not perform well. If there is no residual autocorrelation model performance is not greatly impacted by changing the autocorrelation in the covariate.

Figure 3.16: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 100.

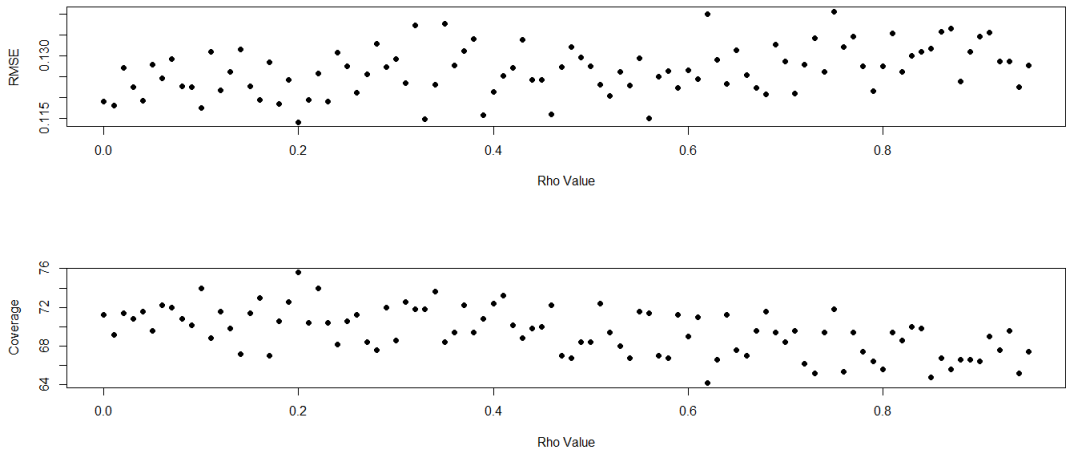


Figure 3.17: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 10.

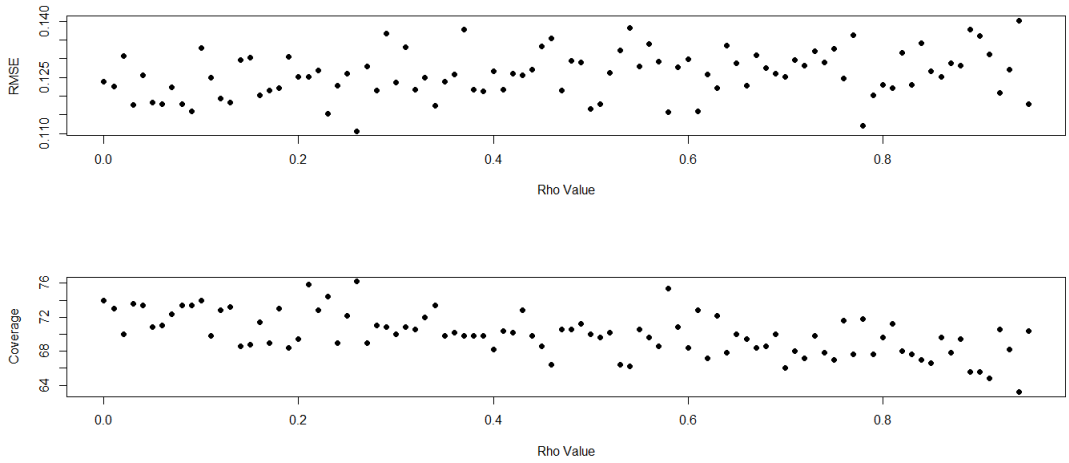
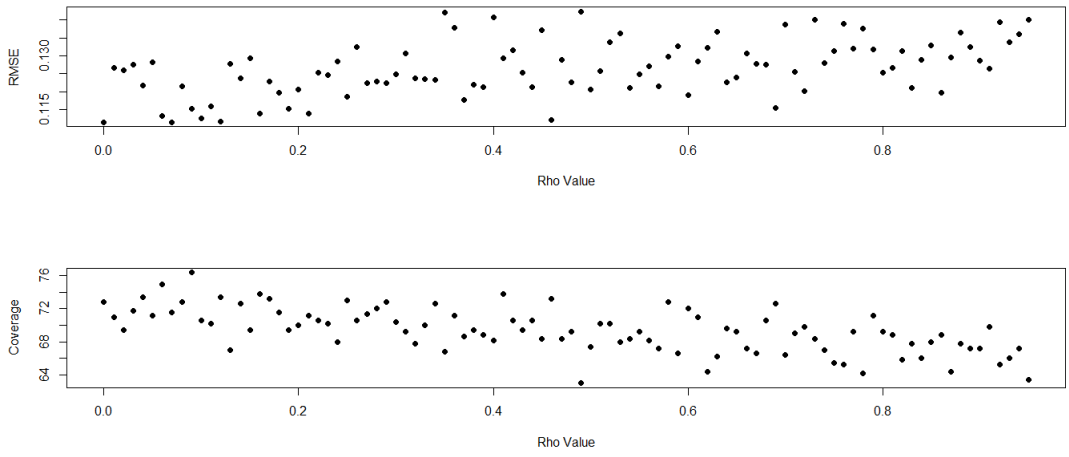


Figure 3.18: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a weakly autocorrelated residual structure and an expected number of disease cases equal to 1000.



3.3.5 Scenario 5 - Varying autocorrelation in the covariate with moderately autocorrelated residual structure

This scenario focuses on how Quasi-Poisson models perform when there is moderately autocorrelated residual structure and the autocorrelation in the covariate is varied. The RMSE (top plot) and coverage probabilities (bottom plot) are shown in Figures 3.19, 3.20 and 3.21 for 100, 10 and 1000 expected cases of a disease in each areal unit. Figure 3.19 show an increase in the RMSE and a decrease in the coverage probabilities as the correlation in the covariate increases toward 0.95. The RMSE ranges from 0.07165, at a minimum, to a maximum of 0.1627, with a mean RMSE value of 0.1008. The coverage probabilities decrease from 72.00%, at a maximum, to a minimum of 38.00%, and the mean coverage is 56.44%. Figure 3.20 shows the RMSE ranges from 0.07742, when the autocorrelation in the covariate is weakest, to 0.1540, when the autocorrelation in the covariate is strongest, with a mean RMSE value of 0.1018. The coverage ranges between a maximum of 71.20% to a minimum of 39.60%, as ρ increases from 0.0 to 0.95, with a mean coverage probability of 58.57%. Figure 3.21 shows the RMSE and coverage when the expected number of disease cases is 1000. The RMSE ranges from a minimum of 0.07013 to a maximum of 0.1513, with a mean RMSE of 0.09978. The coverage decreases, as the autocorrelation in the covariate increases, with a minimum coverage of 37.80%, a maximum coverage of 72.60% and a mean coverage probability of 56.44%.

Figures 3.19, 3.20 and 3.21 show the same trend in RMSE and coverage, specifically an increase in RMSE and a decrease in coverage as ρ increases. The range of RMSE and coverage probabilities are similar for the three considered expected numbers of disease cases, therefore there does not seem to be an effect on results from changing the disease prevalence. The Quasi-Poisson

model performs poorly in terms of coverage as the maximum coverage probabilities found are around 72.00%, which is low compared to the nominal value of 95%. The models perform best, in terms of coverage, when the correlation in the covariate is weakest.

The results found by changing the expected number of disease cases from this scenario are similar to that of scenario four, where the residual spatial autocorrelation is weak, as both scenarios find little impact on model performance from changing the expected numbers of disease cases across the region. However, the effects of changing ρ are very different as there is a stronger trend in this scenario where there is moderate autocorrelation in the residual structure. The coverages show a wider range and have a lower mean value in scenario five, where there is moderate autocorrelation in the residual structure, compared to scenario four.

Figure 3.19: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of disease cases equal to 100.

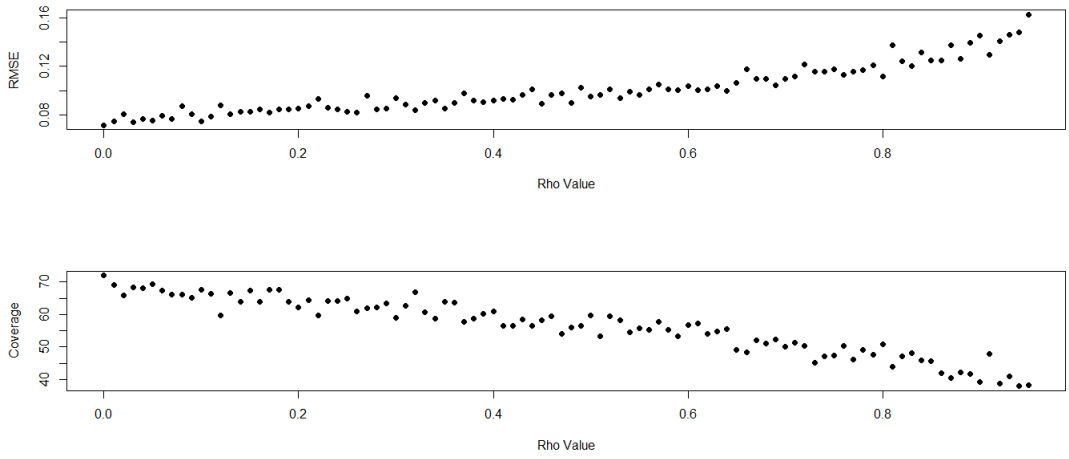


Figure 3.20: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of diseases cases equal to 10.

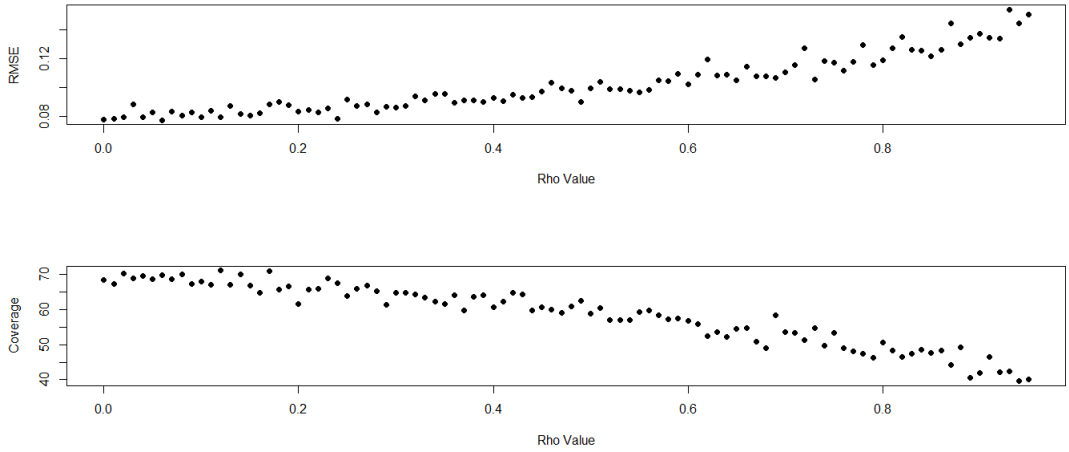
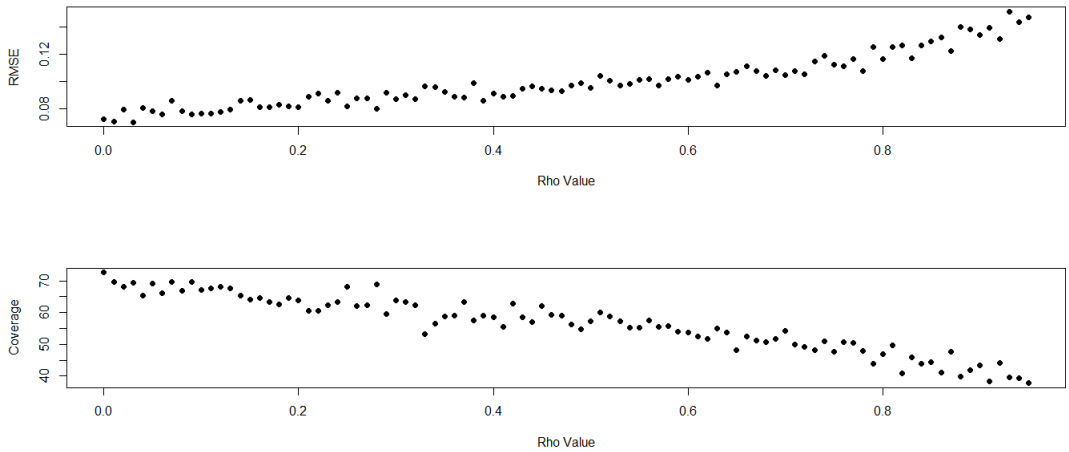


Figure 3.21: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a moderately autocorrelated residual structure and an expected number of disease cases equal to 1000.



3.3.6 Scenario 6 - Varying autocorrelation in the covariate with strongly autocorrelated residual structure

This scenario focuses on how Quasi-Poisson models perform when there is strongly autocorrelated residual structure and the autocorrelation in the covariate is varied. The RMSE (top plot) and coverage probabilities (bottom plot) are shown in Figures 3.22, 3.23 and 3.24 for 100, 10 and 1000 expected cases of a disease across the study region. Figure 3.22 shows a range of RMSE values between 0.04144, when the autocorrelation in the covariate is weakest, and 0.09508, when the autocorrelation in the covariate is strongest, with a mean RMSE of 0.05518. The coverage shows a decreasing trend from 68.80% to 34.20% as ρ increases, with a mean coverage probability of 55.47%. Figure 3.23 shows the RMSE and coverage when the expected number of disease cases is 10. The RMSE shows an increasing trend as ρ increases and ranges between 0.04100 and 0.09189, with a mean RMSE of 0.05703. The coverage decreases as the residual spatial autocorrelation increases, showing a range of values between 43.00% and 74.60%, with a mean coverage of 62.09%. Figure 3.24 shows the results when the expected number of disease cases is 1000. The RMSE shows a range of values between 0.03795 and 0.08662, with a mean RMSE of 0.05501. The coverage ranges between 32.40% and 69.60%, with a mean coverage of 54.17%.

This scenario also shows an increase in the RMSE and decrease in the coverage as ρ increases from 0.0 to 0.95. The Figures, 3.22, 3.23 and 3.24, suggest that changing the disease prevalence has little impact on the results. The Quasi-Poisson model performs poorly in terms of the coverage as the maximum coverage, 74.60%, is below the 95% value. These results are similar to the other scenarios where the autocorrelation in the covariate is varied with a specified strength of autocorrelation in the residual structure. Al-

though, the range of coverage values and the mean coverage probabilities are lower than that in scenarios four and five.

Figure 3.22: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure and an expected number of disease cases equal to 100.

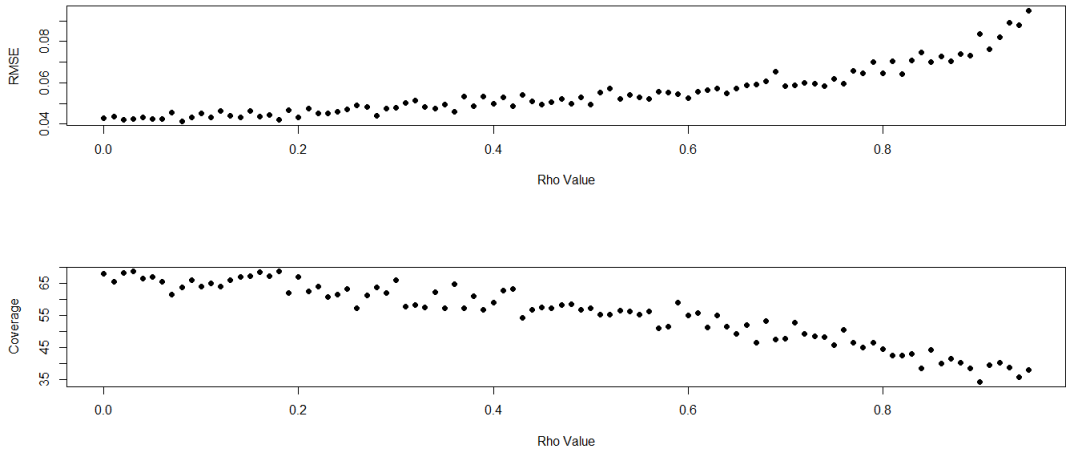


Figure 3.23: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure and an expected number of disease cases equal to 10.

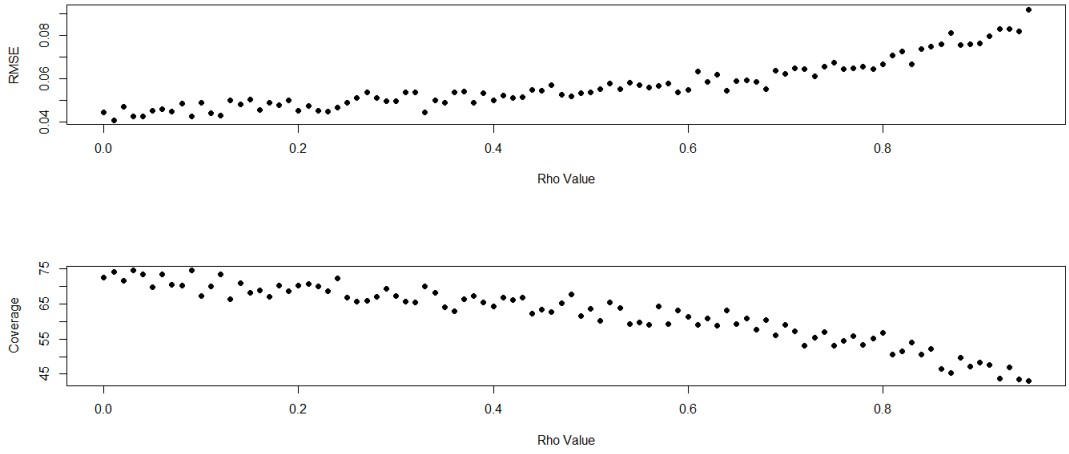
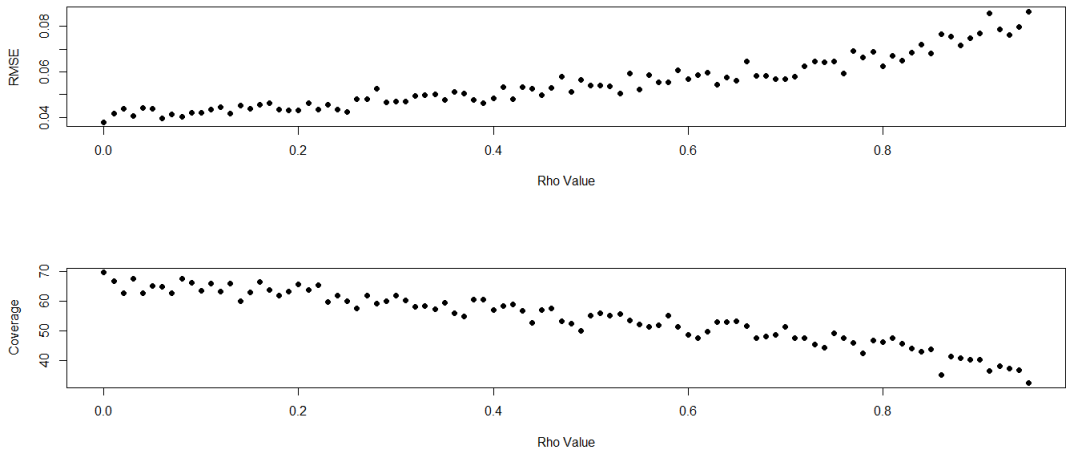


Figure 3.24: RMSE and Coverage for β_1 when the autocorrelation in the covariate is varied with a strongly autocorrelated residual structure with an expected number of disease cases equal to 1000.



3.4 Conclusion

Scenario 1 uses a covariate which is uncorrelated in space and the effect of varying levels of residual spatial autocorrelation is considered. This results from this scenario show that there is no effect from residual spatial autocorrelation when the covariate is uncorrelated, and there is also no effect from changing the disease prevalence. In scenarios 2 and 3 there is an adverse affect due to spatial autocorrelation in the residuals on model performance when there is autocorrelation in the covariate. Scenarios 4, 5 and 6 look at the impact on model performance when the autocorrelation in the covariate is varied with a specified strength of autocorrelation in the residual structure. These three scenarios show an increase in RMSE, and decrease in coverage, as the autocorrelation in the covariate increases, and there seems to be little effect from changing the expected number of disease cases across the region.

The results shown from the six scenarios suggest that spatial correlation

is not a problem in itself. However, spatial correlation is a problem if both the covariate and the residuals are spatially correlated. This could be due to potential collinearity between $\underline{\boldsymbol{x}}$, the covariate, and the random effects $\underline{\boldsymbol{\phi}}$. As the correlation in both $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{\phi}}$ increases at the same time the results get worse, but if correlation in only one of $\underline{\boldsymbol{x}}$ or $\underline{\boldsymbol{\phi}}$ increases whilst the other is independent, there is no drop in performance. As the size of $\underline{\boldsymbol{\phi}}$ gets larger, the model performance gets worse. The Quasi-Poisson model performs poorly in terms of coverage unless $\underline{\boldsymbol{\phi}}$ is very small. Three values for the expected number of disease cases across the region are considered for each scenario, specifically 10, 100 and 1000, and there is little impact on the results found when comparing between the levels of disease prevalence.

Chapter 4

Comparing The Effects of Ignoring and Accounting For Residual Spatial Autocorrelation on Fixed Effect Estimates.

4.1 Introduction

This chapter seeks to compare four different models, namely the Quasi-Poisson generalised linear model, the Poisson log linear model with random effects modelled by a Leroux CAR prior, the Sparse SGLMM proposed by Hughes and Haran (2013) and the Localised Conditional Autoregressive Model (LCAR) proposed by Lee et al. (2014), for a range of circumstances where there is autocorrelation in the covariate and residuals at different levels. The comparisons within this chapter are based upon 500 simulated data sets, where the spatial area of interest is a square 20×20 grid as in the previous chapter. The root mean square error and coverage probabilities of the estimated covariate effects are considered to summarise the model qual-

ity. The Quasi-Poisson model ignores the residual spatial autocorrelation therefore the results found may be inappropriate, and it is of interest to see whether accounting for the spatial autocorrelation shows improved results. The three models achieve this in different ways. The random effects with a Leroux CAR prior is the standard approach to allowing for residual spatial autocorrelation, and assumes the random effects exhibit a single global level of spatial smoothness. However, there is potential for collinearity between such globally spatially smooth random effects and any covariate in the model that is globally spatially smooth. The remaining two models overcome this problem. The model by Hughes and Haran (2013) forces the random effect component to be orthogonal to the covariates, thus preventing this collinearity from occurring. In contrast, the LCAR model proposed by Lee et al. (2014) allows for localised spatial smoothing, thus being less restrictive in the spatial autocorrelation structures that can be estimated. These localised structures are elicited from earlier data on disease risk after the covariate effects have been removed, and thus should not be collinear to the covariates in the model.

4.2 Data Generation and Study Design

The data generation uses the same specification as the previous chapter, and full details are given in Section 3.2. As in the previous chapter, the study region is a square spatial grid of dimension 20×20, yielding $n = 400$ areas in total, from which the distance matrix, D, and neighbourhood matrix, W, are calculated. The neighbourhood matrix W is a binary $n \times n$ matrix where $w_{ij} = 1$ if areas i and j share a common border and $w_{ij} = 0$ otherwise. D is an $n \times n$ distance matrix, where d_{ij} is the Euclidean distance between areas i and j. A total of 500 simulated data sets are generated under a number of different scenarios in this study, to ensure the results are robust to random chance. We consider two different scenarios for the expected number

of cases, the expected number of cases in each area is fixed at either 10 or 100. For each simulated data set the covariate, $\underline{\boldsymbol{x}} = (x_1, \ldots, x_n)$, and the random effects $\underline{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_n)$ are regenerated so that the results are not affected by the particular realisation chosen. The covariate and random effects are each generated from multivariate Gaussian distributions, where the mean is equal to zero and the variance matrix induces a range of spatial correlation structures, ranging from independence to strong spatial correlation. The generation uses the same specification as the previous chapter, and full details are given there.

4.3 Results

Five hundred data sets are generated under a number of different scenarios, which differ in the amount of autocorrelation in the residuals and the covariate. Autocorrelation in the covariate is controlled by changing the value of ρ and in the random effects by changing the value of ν. In this study ρ takes values of 0.0, 0.5 or 0.95 which corresponds to independence, moderate or strong autocorrelation. In addition, ν takes values of 0.1, 1.4 or 30 and these values of ν correspond to a mean autocorrelation of 0.95, 0.50 or 0.007 when calculated by the mean of Σ, where $\Sigma = \exp(-\nu D)$. Additionally, the random effects are scaled by 1 and 0.1, which in the latter case reduces their size by a factor of 10. To compare how the models perform under different disease prevalences, expected numbers of disease cases fixed at 100 and 10 are considered. The results are summarised by showing RMSE (Coverage), for each of the combinations of values, in Tables 4.1 to 4.4.

4.3.1 Small Impact Random Effects

Tables 4.1 and 4.2 present the results for expected disease counts of 100 and 10, respectively, when Var$[\phi]$ is adjusted to be 0.1. When there are 100 expected disease cases, the Leroux model produces results which are almost perfect over the different combinations of the ρ and ν values, as the coverages are found in the 95% to 98% range. The Quasi-Poisson model produces coverage probabilities that are mostly in the 70% to 85% range, with some lower coverages of 46.8% and 47.8% when both the random effects and the covariate are correlated ($\rho = 0.95$ and $\nu = 0.1, 1.4$). The Sparse SGLMM performs slightly better than the Quasi-Poisson model, with most coverages between 80% and 95%, however lower coverages are found when ρ is 0.95 combined with ν values of 0.1 or 1.4. The LCAR model produces coverage probabilities that range between 93% and 98%, which is similar to the results for the Leroux model. The RMSE values are lowest for the Leroux model, followed by the LCAR, then the Sparse SGLMM then the Quasi-Poisson model. The RMSE values for the Leroux model, LCAR and Sparse SGLMM are fairly close in each of the combinations of ρ and ν values, suggesting that correctly accounting for spatial autocorrelation does lead to improved estimation performance. Reducing the expected number of disease cases to 10, allows us to examine how well each of the models perform with a rare disease, and the results are presented in Table 4.2. The coverage probabilities for the Sparse SGLMM increase and range between 87% and 96%. The Leroux model continues to perform well with coverages between 92% and 97% and the Quasi-Poisson model shows coverages between 76% and 93%. The LCAR model performs well, with coverages between 92% and 97%. The RMSE values are lowest for the Leroux model, followed by the LCAR, then the Sparse SGLMM then the Quasi-Poisson model, as in the case with 100 expected disease cases.

4.3.2 Large Impact Random Effects

Tables 4.3 and 4.4 present the results for expected disease counts of 100 and 10, respectively, when Var$[\phi]$ is 1. This scenario is a harder case for all models, as the magnitude of the random effects is now 10 times larger than in the previous subsection. Thus there is a large amount of overdispersion in the data, which may correspond to a large number of unmeasured confounders which affect the response and may be spatially autocorrelated. The results for a disease with 100 expected cases show that the Leoux and LCAR models are the best in terms of coverage as the coverage probabilities range between 74% and 97% for both models. Although most of the coverage probabilities range between 91% and 97%. There are three cases where the Leroux and LCAR models show a poorer performance with coverages of 77.6%, 80.2% and 74.4% for the Leroux model and 77.2% , 75.8% and 73.6% for the LCAR model, and these occur when there is weak autocorrelation in the random effects ($\nu = 30$). The Sparse SGLMM shows coverage probabilities that range between 24% and 47%, suggesting that it cannot cope with the increased amount of unmeasured confounding/overdispersion in the data. The Quasi-Poisson model mostly shows coverages between 63% and 83%, although there are two cases where ρ is 0.95 and ν is either 0.1 or 1.4 when the coverages are lower (38% and 46.2%). In this scenario, the Sparse SGLMM shows a poorer performance than the Quasi-Poisson model in terms of coverage, although the RMSE is lower. The RMSE values for the Quasi-Poisson model are higher than the other methods, the RMSE for Leroux and LCAR are similar and the lowest although the RMSE values for the Sparse SGLMM are generally close to these values. With an expected number of disease cases of 10, the Leroux model continues to perform well with coverages between 96% and 99% in all combinations of ρ and ν values. The LCAR model also continues to perform well, with coverages between 91% and 99%. The Quasi-Poisson model generally shows coverage probabilities around 65% to 80%, except in cases where the autocorrelation in the covariate and resid-

ual are both strong, specifically $\rho = 0.95$ combined with a ν value of either 0.1 or 1.4. The coverage probabilities for the Sparse SGLMM range between 29% and 59%. The RMSE is the lowest for the Leroux model followed by the LCAR, then the Sparse SGLMM and the Quasi-Poisson model. The RMSE values for the Leroux and LCAR are close in all combinations of ν and ρ.

Table 4.1: RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 100 expected disease cases and $\text{Var}[\phi] = 0.1$.

ρ	ν	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
0.0	0.1	0.01105 (73.0)	0.005526 (96.4)	0.006489 (93.6)	0.005543 (95.4)
0.0	1.4	0.01125 (74.8)	0.005673 (97.4)	0.006816 (93.6)	0.005637 (98.4)
0.0	30	0.01073 (76.4)	0.007948 (95.6)	0.009199 (82.4)	0.007970 (95.8)
0.5	0.1	0.01614 (72.0)	0.008212 (95.2)	0.01111 (85.8)	0.008164 (96.0)
0.5	1.4	0.01593 (72.0)	0.008205 (96.2)	0.01093 (84.8)	0.008147 (96.6)
0.5	30	0.01328 (82.6)	0.01089 (95.4)	0.01209 (82.8)	0.01080 (93.6)
0.95	0.1	0.02689 (46.8)	0.008988 (96.8)	0.01665 (71.2)	0.008881 (97.8)
0.95	1.4	0.02522 (47.8)	0.01011 (95.6)	0.01673 (72.8)	0.009893 (95.6)
0.95	30	0.01382 (77.6)	0.01128 (95.0)	0.01314 (81.2)	0.01121 (94.8)

Table 4.2: RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 10 expected disease cases and Var$[\phi] = 0.1$.

ρ	ν	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
0.0	0.1	0.01644 (88.2)	0.01403 (96.6)	0.01499 (94.2)	0.01400 (96.8)
0.0	1.4	0.01614 (90.6)	0.01429 (96.0)	0.01474 (95.2)	0.01421 (95.4)
0.0	30	0.01707 (88.4)	0.01589 (93.6)	0.01663 (92.6)	0.01592 (94.4)
0.5	0.1	0.02494 (87.0)	0.02187 (94.6)	0.02297 (93.6)	0.02159 (94.4)
0.5	1.4	0.02336 (91.4)	0.02107 (95.6)	0.02218 (93.8)	0.02114 (95.8)
0.5	30	0.02187 (92.8)	0.02142 (94.0)	0.02227 (94.2)	0.02138 (94.2)
0.95	0.1	0.03149 (76.4)	0.02502 (92.0)	0.02773 (87.8)	0.02435 (92.6)
0.95	1.4	0.03002 (77.4)	0.02479 (92.6)	0.02647 (92.2)	0.02471 (92.6)
0.95	30	0.02183 (90.8)	0.02131 (94.6)	0.02335 (92.6)	0.02151 (94.8)

Table 4.3: RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 100 expected disease cases and Var$[\phi] = 1$.

ρ	ν	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
0.0	0.1	0.1072 (71.2)	0.02561 (96.6)	0.05857 (47.2)	0.02496 (96.8)
0.0	1.4	0.1122 (71.0)	0.02909 (96.6)	0.05882 (44.4)	0.02775 (97.2)
0.0	30	0.1136 (74.0)	0.08449 (77.6)	0.09435 (24.0)	0.06922 (77.2)
0.5	0.1	0.1677 (63.4)	0.03668 (95.6)	0.1001 (35.8)	0.03659 (96.6)
0.5	1.4	0.1529 (70.6)	0.04285 (94.2)	0.09967 (36.4)	0.04103 (93.8)
0.5	30	0.1319 (83.0)	0.1049 (80.2)	0.1192 (24.0)	0.09424 (75.8)
0.95	0.1	0.2548 (38.0)	0.04952 (93.6)	0.1729 (25.8)	0.04615 (92.0)
0.95	1.4	0.2331 (46.2)	0.05412 (91.4)	0.1512 (33.2)	0.05372 (89.2)
0.95	30	0.1443 (70.4)	0.1139 (74.4)	0.1317 (24.0)	0.09968 (73.6)

Table 4.4: RMSE (Coverage) for Quasi-Poisson, Leroux, Sparse SGLMM and LCAR with 10 expected disease cases and $\text{Var}[\phi] = 1$.

ρ	ν	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
0.0	0.1	0.1269 (68.4)	0.02637 (99.4)	0.06689 (56.6)	0.02757 (98.8)
0.0	1.4	0.1263 (71.0)	0.02879 (99.8)	0.06561 (59.2)	0.03227 (99.0)
0.0	30	0.1242 (69.4)	0.05985 (98.2)	0.1025 (32.0)	0.06550 (92.8)
0.5	0.1	0.1639 (65.6)	0.04122 (99.2)	0.1037 (46.2)	0.04128 (97.2)
0.5	1.4	0.1567 (69.4)	0.04342 (98.6)	0.1018 (47.0)	0.04727 (97.0)
0.5	30	0.1435 (79.0)	0.08216 (98.0)	0.1223 (35.2)	0.08606 (92.2)
0.95	0.1	0.2657 (42.8)	0.05263 (97.2)	0.1892 (29.8)	0.05322 (95.6)
0.95	1.4	0.2351 (46.4)	0.05577 (96.8)	0.1567 (34.2)	0.06183 (94.8)
0.95	30	0.1394 (75.2)	0.09627 (96.6)	0.1238 (36.2)	0.09440 (91.6)

4.4 Conclusion

Ignoring spatial autocorrelation, using a Quasi-Poisson model, results in low coverage probabilities and RMSE values which are fairly large. The Quasi-Poisson model shows some consistency in terms of coverage, although it is outperformed in terms of coverage by the Leroux model in almost all cases and also the Sparse SGLMM when $\text{Var}[\phi] = 0.1$. However, as the Quasi-Poisson model simply ignores the spatial structure of the data it should be expected that the results for methods which account for this structure perform better.

The Leroux and LCAR models perform the best overall, in terms of both RMSE and coverage, showing very similar results across all the scenarios considered. The Leroux model shows coverage probabilities which are mostly close to the 95% level and low RMSE. There are three occassions where this model does not perform as well, these all occur when the expected number of

disease cases is 100 and Var$[\phi]$ is 1, in combination with $\nu = 30$ (no residual autocorrelation). The coverages produced in these cases are 77.6% ($\rho = 0.0$), 80.2% ($\rho = 0.5$) and 74.4% ($\rho = 0.95$). The LCAR model shows a similar performance to the Leroux model, in terms of both RMSE and coverage probabilities. The coverage probabilities are mostly close to the 95% level. The RMSE values are all similar to the values found by the Leroux model. The LCAR also shows three occassions where this model does not perform as well, when the expected number of disease cases is 100 and Var$[\phi]$ is 1, in combination with no residual autocorrelation ($\nu = 30$). The coverages produced in these cases are 77.2% ($\rho = 0.0$), 75.8% ($\rho = 0.5$) and 73.6% ($\rho = 0.95$).

The Sparse SGLMM can perform well with coverages around the 95% level as well as poorly with coverages that can be as low as 24%. The Sparse SGLMM shows reasonable results if Var$[\phi] = 0.1$ and the expected number of disease cases is 10, however the RMSE and coverage probabilities are still worse than found using the Leroux model. Otherwise, the coverage found with the Sparse SGLMM is very bad ranging between 24% and 47% (when Var$[\phi] = 1$, expected disease cases of 100) and ranging between 29% and 59% (when Var$[\phi] = 1$, expected 10 disease cases).

There is the potential for collinearity between the covariate and unmeasured structure when both are spatially autocorrelated ($\nu = 0.1, 1.4$ and $\rho = 0.5, 0.95$). This reduces the coverage and increases the RMSE for the Quasi-Poisson model especially, and there are cases where this occurs with the Sparse SGLMM too. A larger Var$[\phi]$ leads to worse results for all models, although the percentage change is lowest for the Leroux and LCAR in both the RMSE and coverage.

For use on real data, the most appropriate choice of model, on the basis

of these results, would be the Leroux model or the LCAR model due to the consistent performance in terms of coverage and low RMSE values. However, the Leroux model is the simpler of the two so it may be more appropriate to use as the LCAR model does not show much of an increase in performance when compared to the Leroux model. If the choice was to be made between only the Quasi-Poisson model and the Sparse SGLMM, the consistency in results of the Quasi-Poisson model may make it seem the most appropriate if nothing was known about the Var$[\phi]$ as there is some idea of how well it would perform in terms of coverage. Whereas the erratic nature of the coverage results from the Sparse SGLMM would make it hard to determine how appropriate the results gained from real data would be.

Chapter 5

An Application to Central Belt Respiratory Health Data.

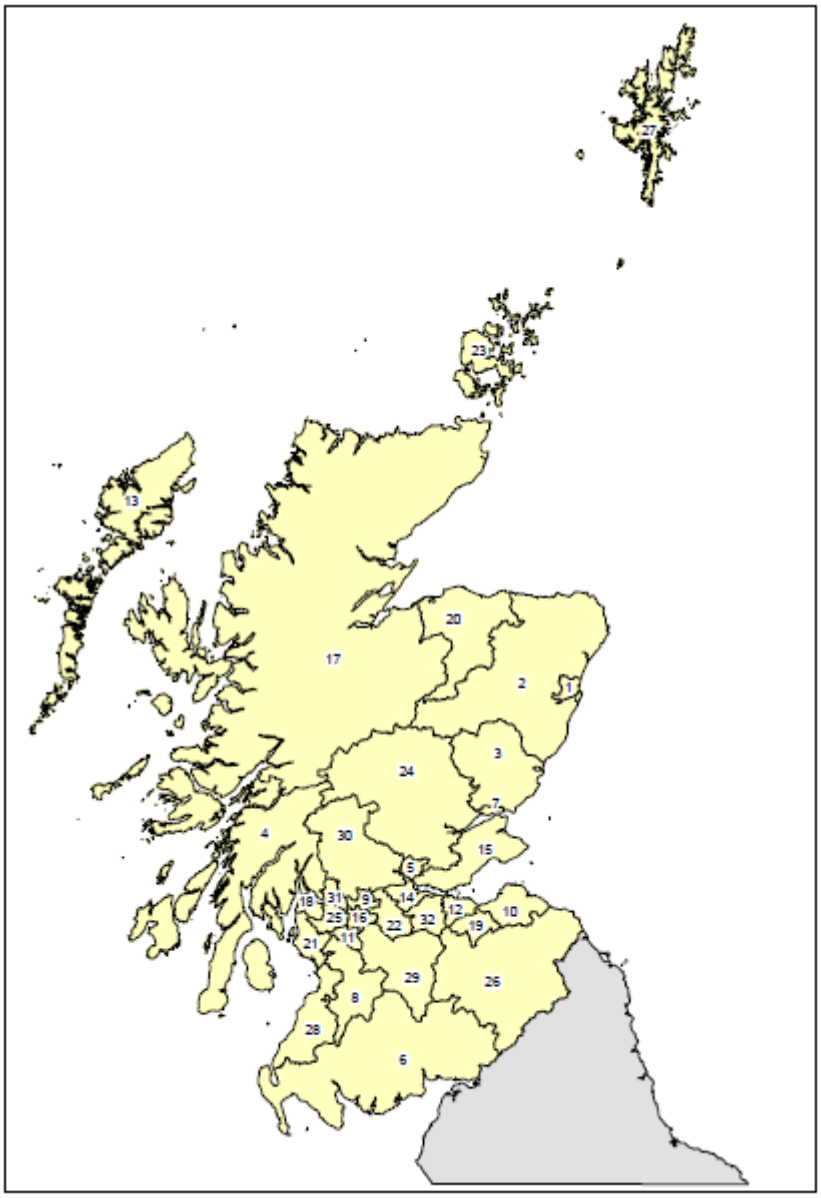
5.1 Introduction

This chapter applies the four methods previously described and used within a simulation study, namely the Quasi-Poisson generalised linear model, the Poisson log linear model with random effects modelled by a Leroux CAR prior, the Sparse SGLMM and the Localised Conditional Autoregressive Model (LCAR), to a real data set of hospital admissions due to respiratory illness. The aim of this chapter is to determine if there is any significant risk associated with increasing pollution levels in the central belt of Scotland on hospital admissions due to respiratory illness whilst accounting for other covariates, such as an indicator of deprivation.

5.2 Data

The data relate to the 545 intermediate geographies comprising the central belt of Scotland.

Figure 5.1: Local Authority Map of Scotland. Source: The Scottish Government (http://www.scotland.gov.uk/Resource/Doc/933/0009386.pdf).



The Scottish central belt is defined to be the local authorities of West and East Dumbartonshire, North Lanarkshire, Midlothian, West and East Lothian, Renfrewshire, East Renfrewshire, Inverclyde, Falkirk, and the cities of Glasgow and Edinburgh. These local authorities correspond to areas 31, 9, 22, 19, 32, 10, 25, 11, 18, 14, 16 and 12 on Figure 5.1. The response

considered is the respiratory health admissions in the central belt during 2010, which corresponds to the international classification of disease (ICD) codes J00 - J99. In 2010 there were 39318 respiratory health admissions in the central belt. The expected number of respiratory hospital admissions for 2010 was computed using external standardisation, based on Scotland rates. These data, along with data on covariates, are sourced from the Scottish Neighbourhood Statistics database. The expected numbers of respiratory health admissions are based on the demographics of the areas population and are calculated by dividing the population living within each area into a number of strata based on their age and sex. The number of people in each stratum is multiplied by the incidence rate for the stratum and the results are summed across all strata for the area to give the expected number of cases for the area. The covariates available are measures of socio-economic deprivation, ethnicity and a measure of how rural or urban an area is. The measure of socio-economic deprivation that will be used throughout this chapter is the percentage of people in each intermediate geography who are defined to be income deprived. The percentage of non-white children is used as a proxy for ethnicity. There is a measure of how rural an intermediate georgraphy is, which is defined on a numerical scale from 1 to 6, where 1 means an area is urban and 6 means an area is rural. Data on the pollutant levels for 2009 are used, and these data are sourced from the Department for Environment, Food and Rural Affairs (DEFRA). The pollutants considered are Nitrogen Dioxide (NO_2, μgm^{-3}), Sulphur Dioxide (SO_2, μgm^{-3}), and PM_{10} (μgm^{-3}) and $PM_{2.5}$ (μgm^{-3}) which are measures of particulate matter in the air less than 10 and 2.5 micrometres in diameter respectively.

Figure 5.2: Map of the SIR for the Central Belt (2010).

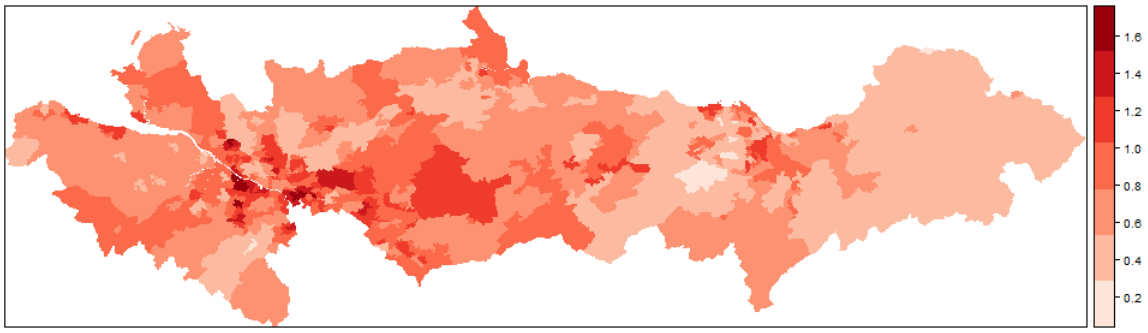


Figure 5.3: Histogram of the SIR for the Central Belt (2010).

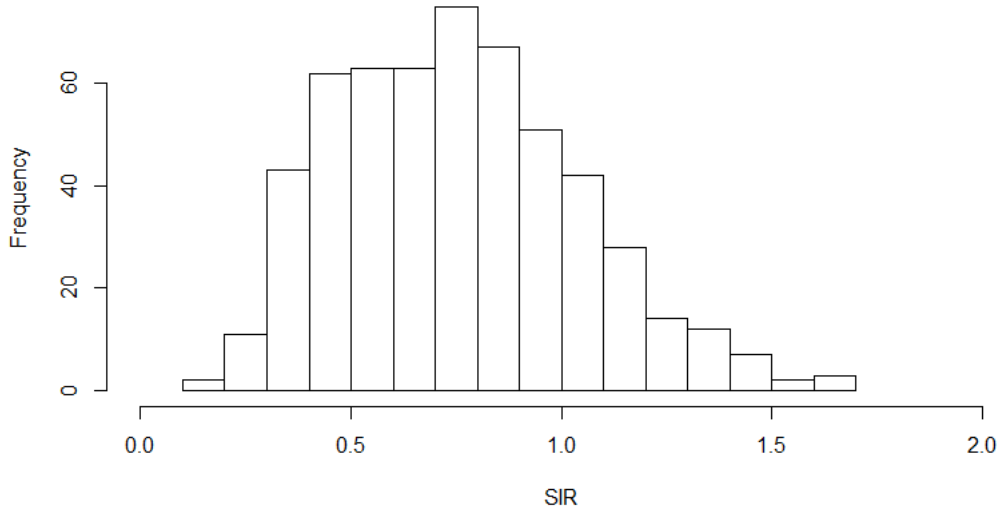


Figure 5.2 shows a spatial map of the central belt showing the SIR values for each area. There is some indication of spatial dependence as in neighbouring regions the SIR values are generally similar. A Moran's I permutation test of the SIR is performed and based on 100000 permutations. The value of the Moran's I statistic is 0.4677, with a corresponding p-value of 1e-05 which is significant at the 5% level. This suggests that there is some spatial associ-

ation within the data. Therefore, analysis should be conducted on the data to try to account for this spatial association. The highest values for the SIR occur in the cities, particularly in Glasgow. Figure 5.3 shows the distribution of the SIR values for the central belt in 2010. The distribution has a mean value of 0.7519, with a median value of 0.7295. The distribution is slightly skewed to the right, with more values at the left of the distribution so there are more areas where the number of respiratory hospital admissions is lower than expected. The SIR values range between 0.1401 and 1.6610, therefore the areas range between 85.99% fewer cases than expected and 66.1% more cases than expected compared to Scotland overall.

Figure 5.4: Boxplot of the Pollutant Concentrations (μgm^{-3}, 2009).

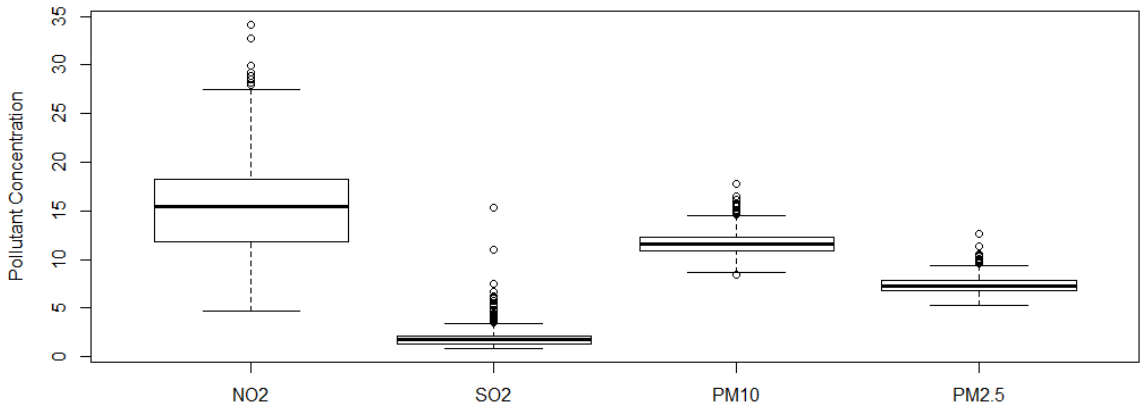


Figure 5.4 shows the concentration of the four pollutants. There is a larger range of values for NO_2, from 4.711 μgm^{-3} to 34.180 μgm^{-3}, whilst the remaining three pollutants show a similar level of variance. The concentration of SO_2 is lowest with a mean value of 1.995 μgm^{-3}, compared to 15.370 μgm^{-3}, 11.630 μgm^{-3} and 7.371 μgm^{-3} for NO_2, PM_{10} and $PM_{2.5}$ respectively. The standard deviations of the pollutants are 5.006427 (NO_2), 1.25221 (SO_2), 1.305974 (PM_{10}) and 0.9537599 ($PM_{2.5}$). The coefficients of

variation are 0.3258 (NO_2), 0.6276 (SO_2), 0.1123 (PM_{10}) and 0.1294 ($PM_{2.5}$). The concentration of PM_{10} is higher than $PM_{2.5}$, however the measure of PM_{10} will include particulate matter that would be measured and included in $PM_{2.5}$.

Figure 5.5: Map of the NO_2 (μgm^{-3}) for the Central Belt (2009).

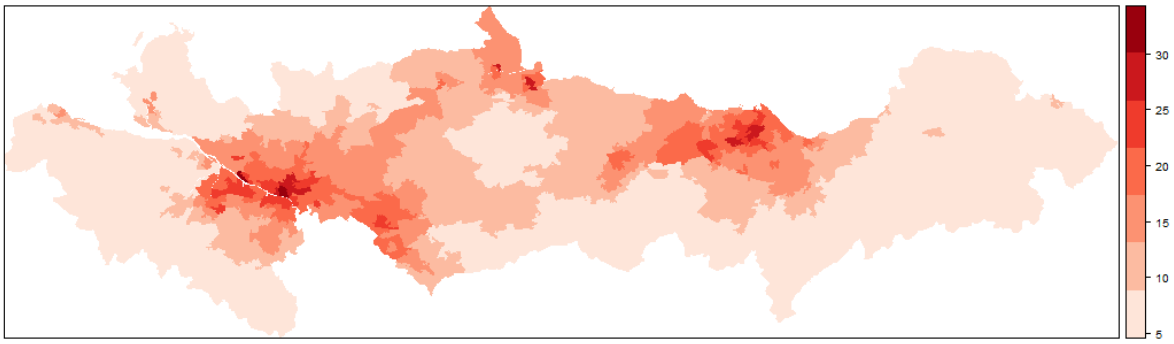


Figure 5.6: Map of the SO_2 (μgm^{-3}) for the Central Belt (2009).

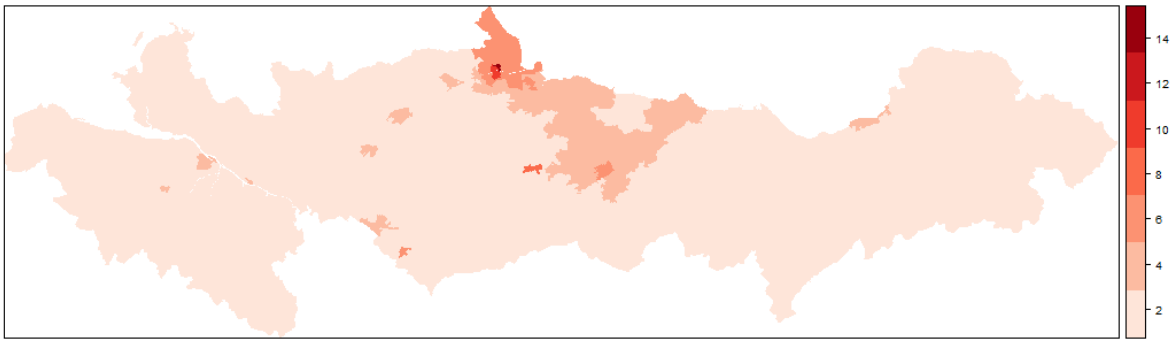


Figure 5.7: Map of the PM_{10} (μgm^{-3}) for the Central Belt (2009).

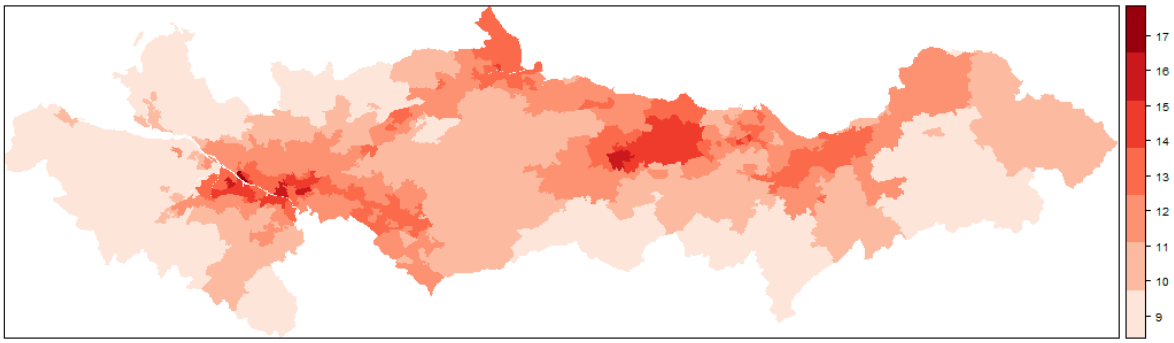
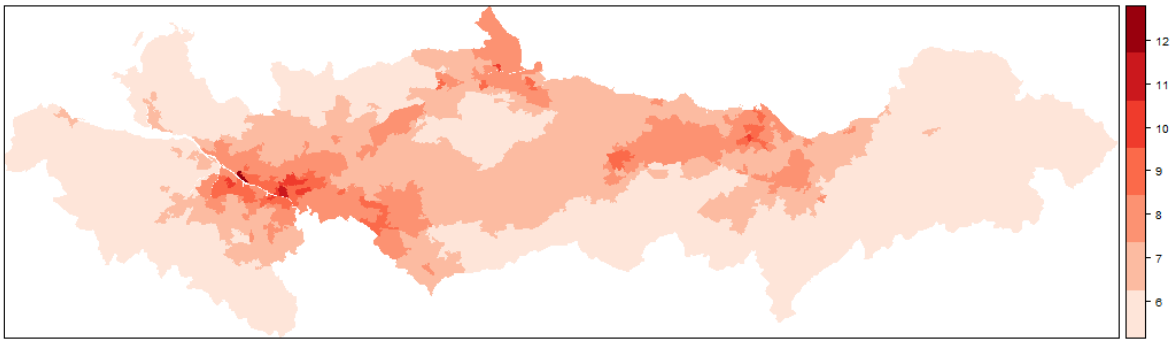


Figure 5.8: Map of the $PM_{2.5}$ (μgm^{-3}) for the Central Belt (2009).



Figures 5.5 to 5.8 show the spatial maps of the four pollutants across the central belt. Figure 5.5 shows that there are higher concentrations of NO_2 in and around the cities of Glasgow and Edinburgh. Figure 5.6 shows that the levels of SO_2 pollution is low in most areas, with slightly higher values in one region of neighbouring areas in the north of the central belt area near Falkirk. This could be due to the Grangemouth refinery located in the Falkirk council area. Figures 5.7 and 5.8 show similar patterns in the concentrations of the pollutants with lower values shown in Figure 5.8. The highest values appear

to occur in the most urban areas such as Glasgow, with rural areas showing lower concentrations of PM_{10} and $PM_{2.5}$.

Figure 5.9: Correlation between the four pollutants.

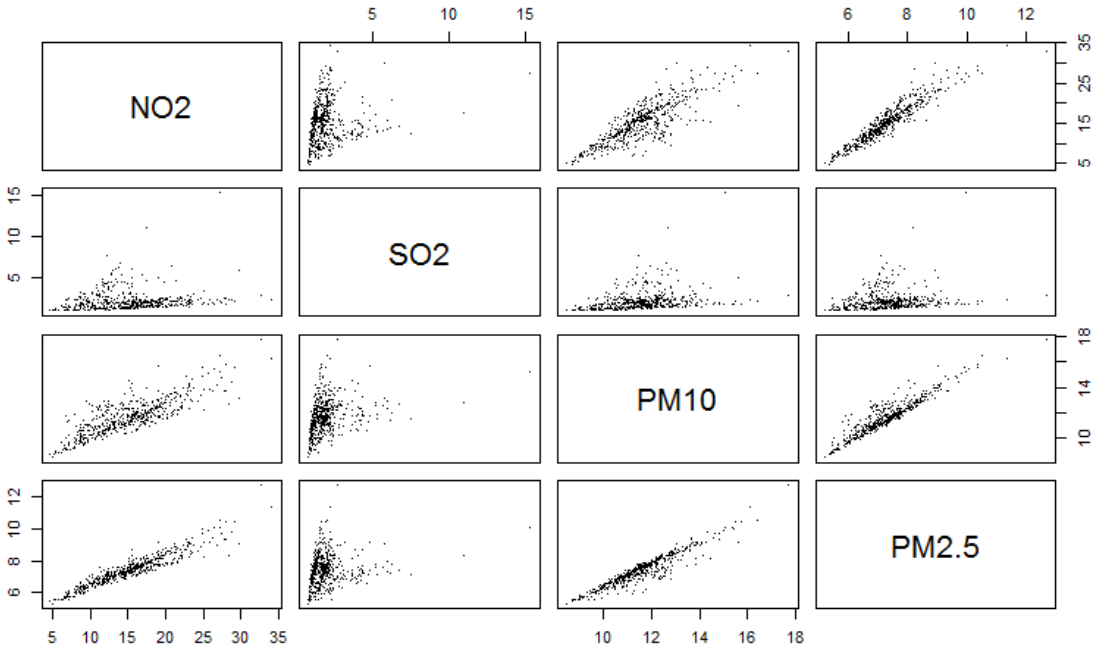


Figure 5.9 shows the relationship between the different pollutants. This indicates that the pollutants show some level of correlation with each other, with PM_{10} and $PM_{2.5}$ showing the strongest correlation ($\rho = 0.9342$) as expected due to the definition of these pollutants as PM_{10} will include all pollution which is measured as $PM_{2.5}$ as well. NO_2 is strongly correlated with PM_{10} ($\rho = 0.8239$) and $PM_{2.5}$ ($\rho = 0.9435$). Due to the strong correlation between the pollutants only one pollutant is included in a model.

Figure 5.10: Map of Income Deprivation for the Central Belt (2009).

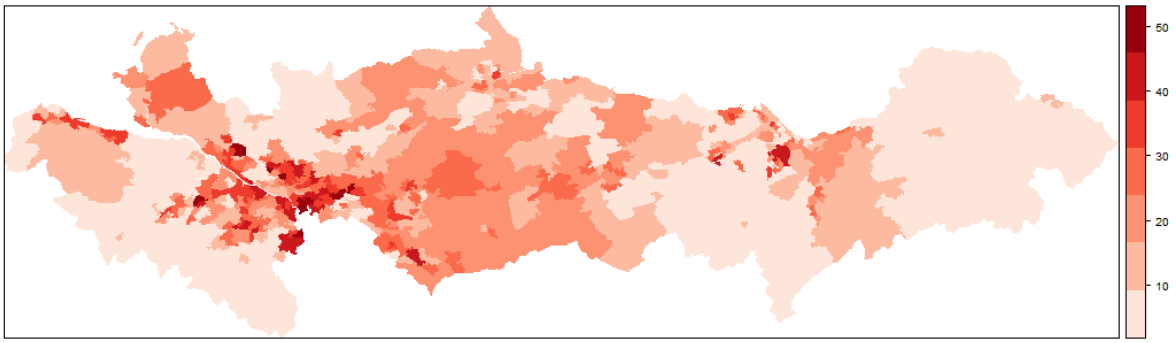


Figure 5.11: Map of Ethnicity for the Central Belt (2009).

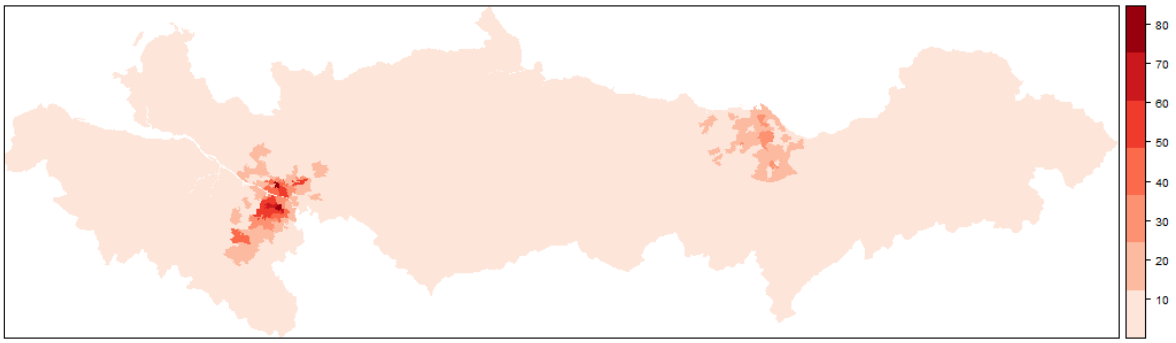
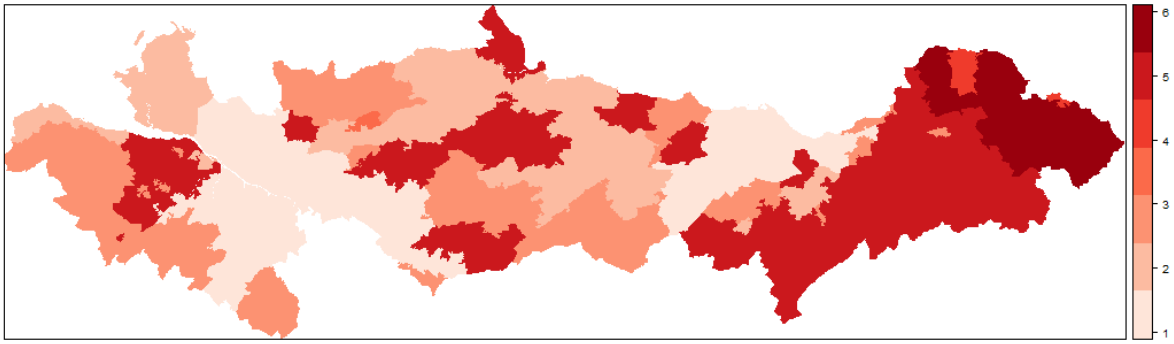


Figure 5.12: Map of Urban for the Central Belt (2008).



The covariates which will be considered in the rest of this chapter are the percentage of people classified as income deprived, the percentage of non-white children and how urban an area is, maps of these covariates are shown in Figures 5.10 to 5.12. Figure 5.10 shows there is a higher percentage of people in the intermediate geographies in the west of the central belt who are defined as income deprived compared to the east. Figure 5.11 indicates that there are higher percentages of non-white children in the cities of Glasgow and Edinburgh with the rest of the central belt showing percentages of non-white children to be less than about 10%. Whereas, in Glasgow the percentage can be around as high as 80%. Figure 5.12 shows how the intermediate geographies are classified into urban and rural areas, with areas corresponding to low values being classified as more urban, and high values relating to the more rural areas. The cities of Glasgow and Edinburgh are the most urban.

5.3 Modelling

Initially, a Quasi-Poisson GLM is fitted for the respiratory hospital admissions in the central belt in 2010 accounting for one pollutant in addition to the three other covariates mentioned previously and shown in figures 5.10 to 5.12.

Table 5.1: Overdispersion Parameter for Quasi-Poisson GLMs.

Pollutant	Overdispersion Parameter
NO_2	3.543157
SO_2	3.535774
PM_{10}	3.457861
$PM_{2.5}$	3.482921

Table 5.1 shows the overdispersion parameters for each of the Quasi-Poisson GLMs fitted. In each pollutant case the overdispersion parameter is about 3.5, which means that the variance of the hospital admissions is about 3.5 times greater than the mean.

Figure 5.13: Residual Plots for Quasi-Poisson GLM Including PM_{10}.

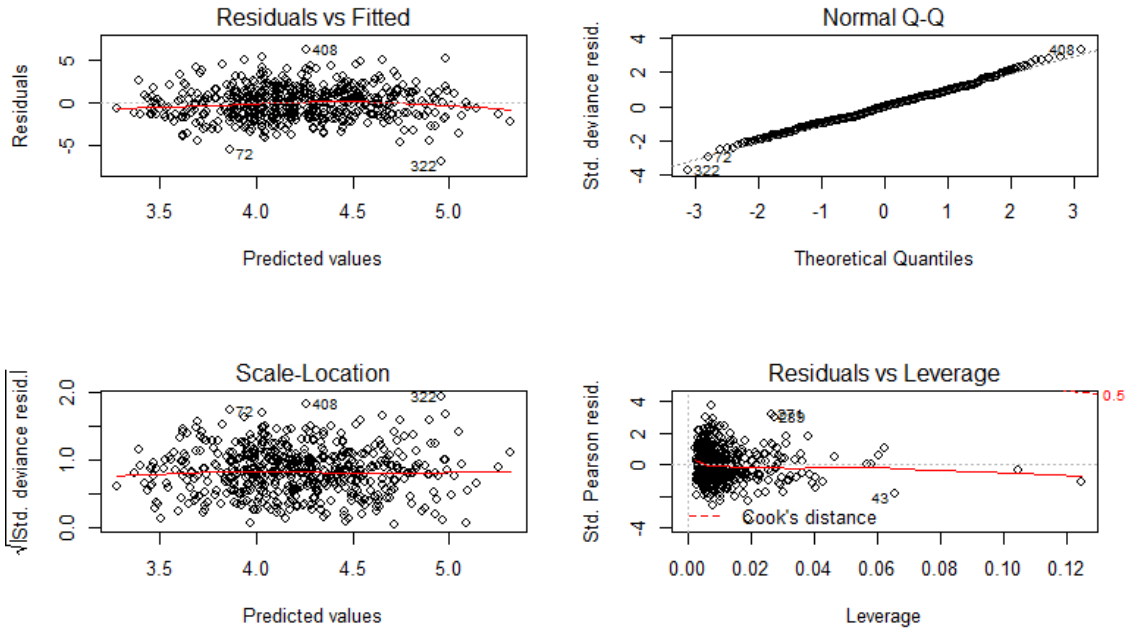


Figure 5.13 shows the residual plots for each Quasi-Poisson GLM fitted in order to assess whether the model assumptions are valid. The assumptions to be checked using figure 5.13, and the residual plots for the other models, are constant variance of the residuals, and that the residuals are normally distributed. The residual versus fitted plots (top left in each of these figures) should show a random scatter of points centered around 0. Figure 5.13 generally shows these qualities and therefore it is suitable to conclude that the assumptions are valid. The Q-Q plot (top right) should show a straight line of points if the residuals are normally distributed. Only some points in this plot deviate from the line with points deviating at both ends. The plots are symmetrical so the assumption does seem appropriate. These plots are produced for each Quasi-Poisson GLM fitted, and the other residual plots are similar to figure 5.13.

Table 5.2: Moran's I for Quasi-Poisson GLMs.

Pollutant	Moran's I Statistic	p-value
NO_2	0.2489	1e-05
SO_2	0.2470	1e-05
PM_{10}	0.2377	1e-05
$PM_{2.5}$	0.2417	1e-05

Moran's I permutation test of the residuals were performed after each model was created, each test is based on 100000 simulations. Table 5.2 shows the values for the Moran's I statistic and the associated p-value. In each pollutant case, the Moran's I statistic is between 0.2377 and 0.2489 which is significant at the 5% significance level. This suggests that there is still some spatial association remaining within the data and further work should be considered to account for this spatial association.

For the three spatial methods, the number of MCMC samples and burn in are fixed and prior data is considered for the LCAR model. The number of MCMC samples generated for the Leroux model is 20000, with the first 10000 samples considered as the burn in period. For the Sparse SGLMM there are 50 Moran eigenvectors used and a maximum number of MCMC samples of 20000. For the LCAR model, data for previous years is needed, for this three years of prior data is used (2007, 2008 and 2009). There are 20000 MCMC samples used with 10000 samples considered as the burn in.

5.4 Results

Table 5.3: Relative Risk (95% Uncertainty Interval) for Pollutants Using Quasi-Poisson, Leroux, Sparse SGLMM and LCAR Models.

Pollutant	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
NO_2	1.0129 (0.9882, 1.0382)	1.0020 (0.9699, 1.0351)	1.0083 (0.9933, 1.0214)	1 (0.9661, 1.0315)
SO_2	1.0173 (0.9967, 1.0379)	1.0127 (0.9895, 1.0438)	1.0126 (0.9994, 1.0245)	1.0108 (0.9851, 1.0396)
PM_{10}	1.0395 (1.0180, 1.0613)	1.0287 (1.0027, 1.0529)	1.0369 (1.0246, 1.0493)	1.0259 (1.0003, 1.0535)
$PM_{2.5}$	1.0363 (1.0128, 1.0602)	1.0226 (0.9962, 1.0531)	1.0322 (1.0198, 1.0457)	1.0239 (0.9935, 1.0518)

Table 5.3 shows the relative risk and the corresponding 95% uncertainty intervals for the relative risk of hospital admission due to respiratory disease for a one standard deviation increase in the pollution concentration. Where the standard deviations for the pollutants are 5.006427 (NO_2), 1.2522 (SO_2), 1.3060 (PM_{10}) and 0.9538 ($PM_{2.5}$). The Quasi-Poisson model suggests that there is a significant risk attached to increased concentrations of PM_{10} and $PM_{2.5}$ as the estimates of the relative risks are 1.0395 and 1.0363 respectively. This means that an increase in the concentration of PM_{10} by 1.3060 μgm^{-3} results in an increased risk of hospital admission due to respiratory illness of 3.95%. Similarly, an increase of 0.9538 μgm^{-3} in the concentration of $PM_{2.5}$ pollution leads to an increased risk of hospital admission due to respiratory illness of 3.63%. The only significant relative risk found using the Leroux model is the risk associated with PM_{10}, where the estimate of the risk is 1.0287. The estimates of the risk using Leroux are similar to those found using the LCAR model. There are two pollutants which show a significant increased risk of hospital admission due to respiratory illness whilst using the Sparse SGLMM. These pollutants are PM_{10} and $PM_{2.5}$ with estimates of the risk of 1.0369 and 1.0322 respectively. This means that there is an increased risk of hospital admission, when the pollutant concentration increases by one standard deviation, of 3.69% (PM_{10}) and 3.22% ($PM_{2.5}$). With the LCAR model, there is a significant increased risk of hospital admission due to res-

piratory illness for increasing the concentration of PM_{10} by one standard deviation. The increased risk of hospital admission due to respiratory illness when increasing the concentration of PM_{10} by one standard deviation found using the LCAR model is 2.59%.

Table 5.4: DIC for each model.

Pollutant	Quasi-Poisson	Leroux	Sparse SGLMM	LCAR
NO_2	–	4197.1721	5119.190	4162.5317
SO_2	–	4195.5085	5009.588	4165.4630
PM_{10}	–	4196.3237	5039.887	4164.7838
$PM_{2.5}$	–	4198.8766	5270.492	4162.7059

Table 5.4 shows the value of the DIC for each model type with the pollutant used in the model. This shows that the LCAR model produces the lowest DIC across all pollutant cases, with the Leroux model showing close results as the DIC is around 30 higher than found with the LCAR model. The lowest DIC value (4162.5317) occurs using the LCAR model with NO_2, and the largest value (5270.492) occurs using the Sparse SGLMM with $PM_{2.5}$. The DIC values for the Leroux model are all similar regardless of the pollutant included in the model, with DIC values between 4195 and 4198. The LCAR model also shows consistent DIC values which range from 4162 to 4165. However, the DIC values found when using the Sparse SGLMM show more variablility as they range between 5009 and 5270, and are around 1000 higher than the DIC found for the LCAR model. Since the DIC value is lowest for the LCAR model, this suggests that this produces the best fit to the central belt data.

5.5 Conclusions

Increasing the concentration of the pollutants by one standard deviation can result in an increased risk of hospital admission due to respiratory illness, in the cases of PM_{10} and $PM_{2.5}$. The results found are all borderline and some show similar estimates for the relative risk but could show different conclusions regarding significance of the risk. However, none of the four methods considered show significant risks associated with increasing the concentration of NO_2 or SO_2 by one standard deviation.

The increased risk of hospital admission due to respiratory illness is likely to be around 2.5-3% for either PM_{10} or $PM_{2.5}$. The relative risk estimates for PM_{10} are 1.0395, 1.0287, 1.0369 and 1.0259 for the Quasi-Poisson GLM, Leroux, Sparse SGLMM and LCAR model respectively. The estimates for the Quasi-Poisson GLM and Sparse SGLMM are close, however these methods were previously shown to be less accurate than the Leroux and LCAR in the simulation study performed in the previous chapter. The increased risk of hospital admission due to respiratory illness associated with increasing the concentration of PM_{10} by one standard deviation found using the LCAR model is lowest at 2.59%. Similar risks are found for $PM_{2.5}$ although there are only significant results found with the Quasi-Poisson GLM and Sparse SGLMM, these relative risks are 1.0363 and 1.0322 respectively.

The DIC is lowest for the LCAR model, in each pollutant case, which suggests that this provides the best fit to the data. The results from the simulation study in the previous chapter show that the LCAR model performs the best in terms of accurate estimation of the fixed effects. This suggests that the relative risk found using this method for PM_{10} is likely to be close to the true value and the uncertainty interval should contain the true value. The relative risk for hospital admission due to respiratory illness for a one

standard deviation (1.305974) increase in the concentration of PM_{10} found using LCAR is 1.0259.

The DIC found for the Leroux CAR prior is close to the values found with LCAR. Since the results found in the previous chapter show that the Leroux and LCAR methods perform similarly, and the relative risks found for PM_{10} are similar it is likely that the true risk is close to these values (1.0287 for Leroux and 1.0259 for LCAR).

The Sparse SGLMM suggests that both PM_{10} and $PM_{2.5}$ affect the risk of hospital admissions due to respiratory illness. The risks found using this method are 1.0369 and 1.0322. These risks are both larger than the risks found using either Leroux or LCAR by around 0.01. This has been shown to be the least accurate in terms of estimation and coverage from the simulation studies, therefore these estimates of the risk may not be the most appropriate.

The Quasi-Poisson GLM shows relative risks which are similar to those found with the Sparse SGLMM and both of these methods show significant results for PM_{10} and $PM_{2.5}$. The Moran's I permutation tests performed after the Quasi-Poisson GLMs were fitted show that spatial autocorrelation is still present, therefore spatial models should be considered as the results may not be appropriate.

Chapter 6

Conclusions.

In this thesis we have considered if there is an improvement in terms of covariate estimation by allowing for residual spatial autocorrelation, and, if allowing for this autocorrelation is better, which method is the best to use. Two simulation studies were performed prior to applying the methods considered to a real data set. The simulation studies allowed the levels of autocorrelation to be controlled, and model performance was evaluated through the RMSE and coverage probability for the estimated fixed effects. A square grid of dimension 20×20 was considered and a total of 500 simulated data sets generated for both simulation studies. The first simulation study focused on the performance of a Quasi-Poisson GLM with six scenarios considered for the autocorrelation. The scenarios allow the autocorrelation in either the covariate or the residuals to be varied whilst holding the other constant at independence, moderate or strong autocorrelation. The second simulation study compares the Quasi-Poisson GLM with models which account for spatial autocorrelation, namely a Leroux CAR prior, the Sparse SGLMM and the Localised Conditional Autoregressive Model. The four methods of modelling were also applied to a data set where the response is the hospital admissions due to respiratory illness in the central belt of Scotland in 2010, with an aim to determine if the level of a pollutant increases the risk of hospital admission due to respiratory illnesses.

The first simulation study suggests that the Quasi-Poisson GLM performs poorly unless the scale of the random effects is very small, as the coverage probabilities are less than 95%. The results also show that as autocorrelation in both the residuals and the covariate increase the model performance gets worse, as the coverage probabilities decrease and the RMSE increases. However, if the autocorrelation in one of the covariate and the residuals increases whilst the other is independent there is no drop in the model performance. This simulation study suggests that spatial autocorrelation has the largest negative impact in fixed effect estimation if both the covariate and the residuals are spatially autocorrelated, which could be due to collinearity between the covariate and the random effects. The three values considered for the expected number of disease cases across the study region suggest that there is little impact on the model performance when changing the level of disease prevalence, as the values found for the RMSE in the three cases are similar in each scenario. For example, in scenario 2 the RMSE ranges between 0.1212 and 0.1744 for 100 expected disease cases, changing the prevalence to 10 the RMSE ranges between 0.1225 and 0.1816 whilst 1000 expected disease cases show a range of RMSE values between 0.1246 and 0.1824. These ranges for RMSE are very similar, and this is also found with the other scenarios.

The second simulation study compares the Quasi-Poisson GLM with three models which account for spatial autocorrelation. The results found show that the Quasi-Poisson GLM results in low coverage probabilities and fairly large RMSE values, with the RMSE ranging between 0.1072 and 0.2657 when the scale of the random effects is 1 for either expected number of disease cases. However, the Quasi-Poisson GLM does show consistency in terms of the coverage probabilities, with most values between 65% and 80%, although some as coverage probabilities as low as 38% to 47% are found. The models which account for the spatial structure of the data outperform the Quasi-Poisson

GLM in almost all cases in terms of coverage probabilities. The best performing methods were the Leroux and the LCAR models, which show very similar results. There are three occassions where these methods do not perform as well, these occur when the expected number of disease cases is 100 and Var$[\phi]$ = 1 in combination with no residual autocorrelation ($\nu = 30$). However, even though these methods do not perform as well in these occasions the methods still outperform the Quasi-Poisson GLM and the Sparse SGLMM. The coverage probabilities in these cases are 77.6%, 80.2% and 74.4% for Leroux and 77.2%, 75.8% and 73.6% for LCAR. Although the other coverages produced by these methods are close to the 95% level. The Leroux and LCAR methods possibly do not perform as well in these cases as there is no residual autocorrelation ($\nu = 30$), therefore it may not be appropriate to try to fit a spatial model in these cases which results in the poorer performance in terms of RMSE and coverage. The Sparse SGLMM can perform well with coverages near the 95% level, however it can also perform poorly as the coverages can be as low as 24%. The results for Sparse SGLMM are reasonable when the expected number of disease cases is 10 and Var$[\phi] = 0.1$. However, the results for the Sparse SGLMM are always worse than the Leroux and the LCAR. The Sparse SGLMM produces narrower uncertainty intervals than the other methods, which impacts on the coverage probabilities found. One possibility for why this method does not perform as well as the Leroux or LCAR is that the known covariates may be correlated to the unmeasured structure. The Sparse SGLMM does not allow for this, whereas the Leroux and LCAR do. There is potential for collinearity between the covariate and the residuals when both are spatially autocorrelated, this reduces the coverage and increases the RMSE. When the scale of the random effects is large the results are worse for all models, although the percentage change is lowest for the Leroux and LCAR in both the RMSE and coverage. This simulation study suggests that the most appropriate models for use on real data are the Leroux or LCAR as they perform well across the scenarios con-

sidered, however as the Leroux is the easiest and the differences between the two methods are small this may be the more appropriate for use on real data.

The application of these methods to a real data set focuses on the risk associated with increasing levels of pollutants on the respiratory illness hospital admissions in the central belt of Scotland. The risk associated with increasing the level of PM_{10} by one standard deviation is likely to be between 2.5% and 3.0%. The risk associated with increasing the levels of $PM_{2.5}$ is likely to be between 1.3% and 4.6% according to the Quasi-Poisson GLM and Sparse SGLMM. The results found for the different pollutants considered are borderline significant or non-significant, which could mean that there is not enough data to find a significant risk. The results show that there are significant risks associated with PM_{10} found using each of the methods considered, however, no significant risks found for either NO_2 or SO_2. The DIC values found for the methods suggest that the LCAR model provides the best fit with the Leroux CAR model performing similarly.

This thesis considers if there is a difference in results of fixed effect estimation from accounting for or ignoring the spatial autocorrelation. The results shown in the previous chapters indicate that it is not appropriate to ignore the spatial autocorrelation, as the coverage probabilities are lower than the spatial methods and are less than 95%, and there are higher RMSE values compared to the methods which account for the residual spatial autocorrelation. Ignoring the spatial autocorrelation is simpler from a modelling perspective, however, the results are inappropriate. Since allowing for the spatial autocorrelation is better than ignoring it, the focus is then on comparing the spatial models to determine which is best. The choice of model for the spatial autocorrelation does make a difference in terms of the fixed effect estimation and three different methods are considered within this thesis. The standard CAR model is shown to work well, despite recent suggestions that

the spatial confounding will adversely affect the fixed effect estimation (Reich et al. (2006), Hughes and Haran (2013)). Whilst the orthogonal method of the Sparse SGLMM may theoretically remove the possibility of confounding, however, it does not work well in practice. The Sparse SGLMM produces erratic results, in terms of coverage probability, unless the scale of the random effects is small. The coverages produced can reach the 95% level when the scale of the random effects is small, however if the scale of the random effects is large then the coverage probabilities found can range between 24% and 60% depending on the number of expected disease cases. The LCAR method also performs well with similar results to the Leroux model across the range of scenarios considered. Both the LCAR and Leroux show coverage probabilities close to 95% in the simulation study described in Chapter 4, and both show similar problems when there is no residual autocorrelation present. The RMSE is lowest for these two methods, with the highest coverage probabilities, and aside from the three occasions with no residual autocorrelation these methods perform consistently well. Overall spatial collinearity is a problem, meaning that appropriate spatial methods are vital for obtaining accurate results. The application of the methods to a real data set confirm recent research which finds that air pollution negatively impacts on human health, with the risk of hospital admissions due to respiratory illness increasing as pollution levels increase.

Future extensions to this work could include examining a wider study region, such as extending the study to Scotland or the United Kingdom. The central belt is the area of Scotland which accounts for the highest density of Scotland's population with the capital city (Edinburgh) and the largest city (Glasgow) included in this region. Therefore the two most urban areas are included, and there are some rural areas included within the central belt but expanding the study region would account for more of the rural communities such as the Scottish highlands. Expanding the area will also include the

rest of Scotland's cities which are smaller in terms of population, however, important in terms of work and industry. For example Aberdeen with the oil industry. Additional work could also involve investigation into why the results found using the LCAR model does not improve much on the results found by the Leroux model in these situations. Further work could include a focus on the temporal autocorrelation to see if the pattern of risk of hospital admission due to respiratory illness changes over time and how the pollution levels change over time. More work could also incorporate the temporal component resulting in an examination of spatio-temporal autocorrelation. Since the models considered here are all spatial models which do not have the ability to incorporate a temporal component, this would require these methods to be developed in such a way to incorporate temporal autocorrelation or the development of new methods.

Bibliography

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society B 36*, 192–225.

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

Clayton, D., L. Bernardinelli, and C. Montomoli (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology 22*(6), 1193–1202.

Dominici, F., R. Peng, M. Bell, L. Pham, A. McDermott, S. Zeger, and J. Samet (2006). Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. *JAMA 295*(10), 1127–1134.

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

Haining, R., J. Law, R. Maheswaran, T. Pearson, and P. Brindley (2007). Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. *Stochastic Environmental Research and Risk Assessnent 21*(5), 501–509.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika 57*, 97–109.

Hughes, J. and X. Cui (2013). *ngspatial: Classes for Spatial Data*. Minneapolis, MN. R package version 1.0-3.

Hughes, J. and M. Haran (2013). Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*, 139–159.

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

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

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

Lee, D. and R. Mitchell (2014). Controlling for localised spatio-temporal autocorrelation in long-term air pollution and health studies. *Statistical Methods in Medical Research*.

Lee, D., A. Rushworth, and S. Sahu (2014). A Bayesian Localized Conditional Autoregressive Model for Estimating the Health Effects of Air Pollution. *Biometrics 70*(2), 419–429.

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

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

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of State Calculations by Fast Computing Machines.

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

Nafstad, P., L. Lund Håheim, T. Wisløff, F. Gram, B. Oftedal, I. Holme, I. Hjermann, and P. Leren (2004). Urban Air Pollution and Mortality in a Cohort of Norwegian Men. *Environmental Health Perspectives 112*, 610–615.

Pope, C. A. (1991). Respiratory Hospital Admissions Associated with PM_{10} Pollution in Utah, Salt Lake, and Cache Valleys. *Archives of Environmental Health: An International Journal 46*(2), 90–97.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reich, B. J., J. S. Hodges, and V. Zadnik (2006). Effects of residual smoothing on the posterior of fixed effects in disease-mapping models. *Biometrics 62*(4), 1197–1206.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, London.

Ruidavets, J., M. Cournot, S. Cassadou, M. Giroux, M. Meybeck, and J. Ferriéres (2005). Ozone Air Pollution is Associated With Acute Myocardial Infarction. *Circulation 11*, 563–569.

Smith, A. and G. Roberts (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society Series B 55*, 3–23.

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