Bayesian Modelling of the Spatial Distribution of Road Accidents

Yilin Liu

A thesis submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Middlesex University Business School

April 2008

PAGE

NUMBERING

AS ORIGINAL

This thesis is dedicated to my parents with love and gratitude.

Acknowledgements

This thesis is the result of four and half years of work whereby many people have assisted and supported in various forms. I would like to have this opportunity to express my gratitude for all of them. I am extremely grateful to Middlesex University for financially supporting the completion of PhD as well as the presentations of my papers in two conferences and to the Research Office in Business School for providing necessary supports during my study.

My first supervisor, David Jarrett, deserves a special mention. He made excellent supervision on the progress of my work and always was available when I needed his advice. He provided many valuable comments on earlier versions of the thesis. He is the one who brought my interests in the areas of road safety and applied statistics. I have learned a lot from him. To my second supervisor Chris Wright, thank you for your support and guidance, and giving thoughtful comments on the draft of thesis. Chris introduced the area of road safety to me and always encouraged me during my study. He also helped me obtain the right to use some data needed for this study from external organisations. To Jeff Evans, thank you for your advice and encouragement during my study. I would also like to thank late Ken Lupton for introducing the Ordnance Survey data and geographical information systems to me.

This study involved the analysis of a large amount of data. I would like to thank Jacqui Bates and Sonal Ahuja of Mott MacDonald Ltd, Anne Patrick of the Ordnance Survey, and Solihull MBC and Coventry City Council for providing access to the SPECTRUM database. The accident data, STATS19, were distributed by UK Data Archive (Crown Copyright) and authored by Department for Transport.

Abstract

This research aims to develop Hierarchical Bayesian models for road accident counts that take account of the spatial dependency in the neighbouring areas or sites. The Poisson log-linear model is extended by introducing a second level of random variation that includes a conditional autoregressive (CAR) component. Both models for accidents at the area level and models for accidents on a road network are developed. Areal models are fitted using data for counties and districts in England covering two different periods and data for wards in the West Midlands region in 2001. Network models are fitted to link data for the M1 motorway and to junction data for the city of Coventry.

Results show that, in most cases, adding a spatial (CAR) component to conventional models produces better estimates of the expected number of accidents in an area or at a site. Signs of the coefficients for explanatory variables, including level of traffic and road characteristics, are consistent with expectation. Levels of the spatial effects in a CAR model reflect the relative influence of the unknown or unmeasurable explanatory variables on the expected number of accidents. Results from models at the local authority level in the 2000s show that spatial effects are positive in London boroughs and are negative in most metropolitan districts. For accidents at the ward level in the West Midlands, the performance of the CAR model is similar to that of the non-CAR model which includes log-normal random effects and metropolitan county effects. For models of accidents on the M1, several links are identified to have positive and fairly large spatial effects. For Coventry junction accidents, the CAR model does not perform better than the non-CAR model. Approaches to including temporal effects in spatial models when data cover two or more periods and jointly modelling different types of accidents are also proposed and examined.

Two applications of the CAR models developed in this research are introduced. The first application is about predicting the number of accidents in a local authority in a new year based on previous years' data. One advantage of using the CAR model is that it produces more precise predictions than the non-CAR model. The second application of the CAR model is a new approach for site ranking. The sites selected by such a criterion are those with high risks caused by some unknown or unmeasured factors (for instance, curvature or gradient of roads) which are spatially correlated. Further on-site investigation will be needed to identify such factors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Every year in the world, road accidents cause injury and death, and result in a serious economic burden. Although in Great Britain the numbers of people killed and seriously injured have been reducing year by year, road safety still remains a serious problem. There were 280,840 casualties and 200,700 reported road accidents involving personal injury in Great Britain in 2005 (Department for Transport, 2006*c*). The total cost-benefit value of prevention of road accidents in 2005 was estimated to be over £18 billion (Department for Transport, 2007*a*).

Against this background, in March 2000 a new road safety strategy 'Tomorrow's Roads—Safer for Everyone' was published by the government (see Department for Transport, 2001). It establishes challenging casualty reduction targets to be achieved by 2010. Compared with the baseline averages for 1994 to 1998, it aims to achieve:

1. a 40% reduction in the number of people killed or seriously injured in road accidents;

2. a 50% reduction in the number of children killed or seriously injured;

3. a 10% reduction in the slight casualty rate.

While good progress has been made towards the government targets, further reductions in casualties are needed. Numbers of casualties are closely related to the numbers of road

accidents that involve injuries. Therefore, any reduction in casualties is associated with a reduction in injury accidents. In order to reduce annual road accidents at the national level, similar targets for reductions in accidents are set at the regional level, for instance for local authorities. In addition, to improve road safety, causes of road accidents and the relationship between numbers of accidents and relevant factors, such as the level of traffic and road geometry, need to be investigated and studied.

As described by Ogden (1996), road traffic may be considered as a system that consists of various components. These components, such as the human, the vehicle and the roads, interact with each other. An accident may be considered as a failure in the system. The UK Department of Transport (Department of Transport, 1986) defines an accident as 'a rare, random and multi-factor event always preceded by a situation in which one or more persons have failed to cope with their environment'. Whether motivated by a humanitarian, public health or economic concern, the analysis of previous accident data is needed in road safety research. Statistical methods have been used in this area for a long time. With the development of the theory of generalized linear models (McCullagh and Nelder, 1989), improved methodologies for analysing road accidents become available. However, researchers face more challenges today. The number of journeys and the volume of traffic on the road become greater and greater especially in the developing countries. The road network structure becomes more complicated. There are more interactions between road users and the road environment. All these situations generate more information and more complex road user behaviour for researchers to cope with.

The approach to modelling of numbers of road accidents has experienced several stages in the development of statistical technique. Many statistical models have been developed to relate the accident count to demographic characteristics, road geometry and traffic characteristics (see, for instance, Jarrett et al., 1989; Miaou, 1994; Milton and Mannering, 1998). The response variable is usually an accident frequency, that is, the total number of accidents of a particular type (for instance, determined by severity) in a wider geographical area (for instance, a local authority) or at a site (for instance, a link or a junction) during a fixed period of time. It is usually assumed to be Poisson distributed.

However, the mean of the Poisson distribution can vary from area to area or from site to site and depends on the characteristics of the area or site, such as the level of traffic and road geometry. In order to relate the accident count to such explanatory variables, conventional approaches usually apply generalized linear models, fitted by maximum likelihood. Maher and Summersgill (1996) give a broad review of the statistical methodology for accident models. The Poisson log-linear model and the negative binomial model are two well-known forms of model in road safety research. They are special instances of generalized linear models. The former does not perform well when the data display *over-dispersion*, that is the residual variance is larger than the fitted Poisson mean. Such over-dispersion can be taken account of by introducing an extra level of random effects that follow a gamma distribution in the Poisson mean. This leads to a negative binomial model that is much used in recent road safety research.

Models of accidents at different spatial levels should be used for different research purposes. In this thesis, models of accidents at the local authority or ward level are called *areal models* and models of accidents at sites in a road network are called *network models*. Areal models can be used to study the relationship between accident frequencies and factors like road conditions and economic development. They can also be used to compare accident frequencies in different administrative areas, such as local authorities, during the same period or to study changes in accident frequencies in an area in different years. See for instance Jarrett et al. (1989); Levine et al. (1995*b*); Miaou et al. (2003). Network models look at road accident frequencies at a more local level (on links or at junctions) and are usually developed to investigate the relationship between accident frequencies and contributory factors such as traffic flow and features of road geometry like road width, number of lanes and type of junction (see, for instance, Layfield et al., 1996; Summersgill et al., 1996, 2001; Walmsley, Summersgill and Payne, 1998). They can also be used for prediction (see, for instance, Greibe, 2003; Mountain et al., 1996; Qin et al., 2003). Moreover, they can be applied to identify high-risk sites and to evaluate the effectiveness of engineering treatments on selected sites (see, for instance, Hauer, 1997; Miaou and Song, 2005; Mountain et al., 1995a).

The performance of areal models and network models with respect to the above mentioned research purposes depends on how precisely the expected accident frequency in an area or at a site can be estimated, which in turn is determined by the explanatory power of the statistical models. In principle, if all the contributory variables for road accidents could be identified and measured, and if the correct form of the model were known, the expected accident frequency would be estimated well. However, in practice, variables like traffic levels cannot be measured precisely and there are other contributory factors that are difficult to measure or even not known. In addition, a 'true' model for accident counts is not known. Therefore, a great challenge in road safety research is to improve statistical models for road accidents by taking account of contributory factors that are not directly measured by the explanatory variables. Under such circumstances, there are several ways to think about the problem.

First, if data for some contributory factors are not available, is it possible to find variables that can be measured and used as proxies for the contributory factors? Sometimes, the answer will be yes. For instance, Bailey and Hewson (2004) and Noland and Quddus (2004) use employment and resident population as proxies for traffic levels. However, proxy variables cannot be found for all the contributory factors. Moreover, one limitation of using the proxies is that they are not best approximations to true explanatory variables and so will introduce measurement error bias (see Maher and Summersgill, 1996).

Another way to account for the unobserved or unmeasurable contributory factors is to consider the information that accident data contain. Accidents can be categorised by features like severity or road class. When multiple response models are developed for different types of accidents (for instance, determined by severity), it is possible that there are some common contributory factors for different types of accidents. When such factors are not included in the model, the correlation in the multiple response variables will not be fully explained by the model. But since the common contributory factors may have similar influence on the expected numbers of accidents of different types, the variation in the response variables due to such factors could be partly explained by introducing some random effects in the Poisson means that model the correlation in the expected numbers

of accidents of different types. Studies that jointly model different types of accidents include Tunaru (2002) and Song et al. (2006), both of which use a Bayesian modelling approach.

A further way of thinking about the effects of the unobserved or unmeasurable contributory factors is to consider the characteristics of accident models. Both areal models and network models of road accidents are a type of spatial model because the observational unit is a location. *Spatial data* are collected and aggregated over space and likely to be spatially correlated. For areal models, factors like the extent of development and urbanization are more likely to be similar in neighbouring areas. For network models, factors such as traffic levels and road geometry (for instance, curvature and gradient) are likely to be spatially correlated for neighbouring sites. When such factors are not measured perfectly or are unmeasurable, variation in the response variables cannot be completely explained. However, since such factors are often spatially correlated, by introducing some appropriate form of spatial random effects in the models, variation in the response variables due to these factors can be partly explained (see Besag, in his contribution to the discussion of McCullagh, 2002). Therefore, better estimates of the expected numbers of accidents in different locations can be achieved.

The spatial correlation in the contributory factors indicates that the means of the Poisson distributions for different areas or sites should also be spatially correlated. However, conventional models with random effects treat these Poisson means as independent. Such models do not take account of any spatial effects. They ignore the possible spatial dependence between different areas or sites, especially between neighbouring areas or sites, and therefore may not fully account for the spatial variation in the response variable—residuals from the model may be spatially autocorrelated especially when not all the spatially correlated contributory factors are included in the models. For areal models, areas that share common boundaries are unlikely to be spatially independent. Moreover, in the context of road accidents, traffic moves on the roads, accidents occur on the roads and adjacent areas always have some common roads passing between them. For network models, the road network itself displays a spatial structure that defines the spatial dependency

among the sites. Both of these features indicate that the spatial independence assumption is not appropriate. In order to remove the spatial correlation in the residuals caused by the incomplete inclusion of spatially correlated contributory factors, one approach is to borrow spatial information implied by a geographical map (for areal models) or a road network (for network models) and introduce spatial effects in the models. Such spatial random effects need to be spatially structured (correlated).

In general, when models include some complicated form of spatial effects, computation is difficult using the frequentist approach. The Bayesian approach facilitates the inclusion of different random effects by formulating them in different layers via a hierarchical structure. Recent progress in Markov Chain Monte Carlo methods and their computer implementation make the computation of Bayesian models more convenient and faster. Studies adopting the Bayesian approach in recent road safety research include Tunaru (2002), MacNab (2003), Miaou et al. (2003), and Bailey and Hewson (2004).

Models with spatial effects are expected to produce better estimates of accident frequencies. This will benefit researchers and engineers in the following ways. First, more reliable conclusions about the reduction of accident frequencies over time can be made. Secondly, better predictions of accident frequencies in the future can be obtained. Moreover, high-risk sites identified using spatial effects, will help safety engineers to find further insights of road network design and urban planning on the occurrence of road accidents.

1.2 Aims of the research

As explained above, for areal models and network models the spatial correlation in the expected numbers of road accidents in neighbouring areas or at neighbouring sites needs to be considered. However, very little work has been done on this aspect. Therefore, the main aim of this research is to develop accident models with spatial effects. A Bayesian modelling approach is adopted. Generally speaking there are two reasons for using this approach. First, the inclusion of spatial and temporal random effects makes the models complicated and difficult to fit by the frequentist approach. Second, Bayesian models

used in disease mapping have been well developed to take account of spatially structured random effects (see Best et al., 2005). However these models need to be modified in order to make them more appropriate for accident data. In this research, both areal models and models for accidents on a road network have been developed.

For areal models, the main objectives are:

- to develop univariate spatial models for accident frequencies that include spatially structured random effects;
- to study the effects of adding spatial effects to the conventional models—this includes the comparison of conventional models and spatial models according to measures of model performance and the results from residual analysis;
- to develop univariate spatio-temporal models for accident frequencies over two or more years that take account of both spatial effects and temporal effects;
- to extend these univariate models to multivariate models that, for example, jointly model accidents of different severities;
- to study the relationship between accident frequencies and variables like traffic volume and road lengths.

For models at the road network level, the objectives are:

- to examine the extent of spatial correlation in the accident data for a road network;
- to develop spatial models for accidents on a road network, considering spatial correlation in neighbouring sites;
- to study the effects of the inclusion of spatial effects.

From the point of view of road safety research, this research aims to give better predictions of the numbers of certain types of accidents in a location in the future, based on the spatial models developed in this thesis. It also aims to provide further methods for site ranking—for instance, ranking sites according to the unobserved spatial effects estimated in the spatial models.

This research will contribute to the research methodology in road safety and provide a better modelling approach for accident data by considering spatial effects. It is expected to benefit the government to make decisions on safety policies, road network design and site selection for engineering treatment.

1.3 Overview of the thesis

Chapters in this thesis are organised as follows. The thesis first introduces the research background and reviews the conventional approaches to modelling accident counts. Based on the examination of several types of spatial model used in other disciplines, it then explains how conventional accident models can be improved and proposes the modelling approach in this research. This is followed by applications that fit models using some real datasets. Finally, it illustrates how models that are developed in this thesis can be used in practice with a summary of the findings and the contribution that the research has made.

Chapter 2 reviews the literature about statistical models for road accidents. Statistical models play an important role in road safety research, so the general aims of road safety research and why the statistical modelling approach is important are explained. In Great Britain, road accident data are recorded in a national road accident collection system and database known as STATS19 (Department for Transport, 2005), which provides information for each recorded accident at a very detailed level. This allows for different types of analysis of accident data. Therefore, the main features of accident data are described. The next part of the chapter reviews conventional statistical models for accident frequencies and some existing methods that analyse the spatial aspects of accident data. Limitations of conventional methods are discussed and how they can be improved are explained. A Bayesian modelling approach is adopted in this study. The remainder of the chapter aims to give some details of Bayesian methods for data analysis, including Bayesian computation, estimation of parameters and model comparison. A simple form of Bayesian model for accident frequencies is used to explain how to specify priors and hyper-priors in the Bayesian context.

As discussed in Chapter 2, the independence assumption on the response variables in

accident models is not appropriate. In order to take account of the spatially dependent relationship in the response variables, a suitable modelling approach is needed. Therefore, Chapter 3 explores some existing modelling approaches for spatial data and suggests appropriate models that can be applied to accident data. Spatial modelling approaches have been extensively developed and studied in the area of disease mapping. Two usual forms of spatial models are the conditional autoregressive (CAR) model and the spatial moving average model. The structure and properties of these models are introduced, in each case followed by a discussion of the possibility of applying such models to road accident data. In addition, this research aims to extend univariate spatial models for road accidents to multivariate models, that jointly model accidents of different types (for instance, fatal, serious and slight accidents), and include temporal effects as well. Therefore, some relevant approaches in disease mapping are reviewed next. A statistic of measuring spatial correlation in the data, namely *Moran's I* is introduced at the end of the chapter.

Chapter 4 and Chapter 5 constitute the methodology part of this thesis. Based on the discussion in Chapter 3 about which modelling approach is appropriate for taking account of spatial effects in accident models, Chapter 4 proposes the modelling approach used in this research. Models are developed in order of increasing complexity. Starting from a univariate model without any random effects, the simplest Poisson log-linear model is extended to more complicated models by including different types of random effects. Choices of the neighbours list and the weighting schemes for the spatial CAR models are explained for areal models and network models respectively. Later, software used to fit the models and general rules for model fitting and checking are described. How Moran's I statistic might be used in a Bayesian framework is discussed in the end of the chapter.

The first section of Chapter 5 explains what types of explanatory variables need to be included in accident models and how they are normally measured. The second section of the chapter introduces details of the data used in this research, including the choice of explanatory variables, the sources of data and how the data are restructured or transformed. For areal models, three datasets were used. Two of them include data for local authorities in England during different periods. One is from 1983 to 1986 and the other is

from 2001 to 2005. Another set of data is for wards in the West Midlands in 2001. Two datasets were used to develop models for accidents on a road network. One is for annual link accidents on a motorway in England from 1999 to 2005 and the other is for accidents at major junctions aggregated for five years in Coventry.

After the introduction of the modelling approach and the data used in this research in Chapters 4 and 5, results of the model fits for areal models and models of accidents on a road network are presented in Chapter 6 and Chapter 7 respectively. The first two datasets are used to fit the models at the local authority level. Both spatial and temporal effects are considered in the models. Multivariate models are fitted using the more recent dataset. Multivariate models with spatial effects are fitted at the ward level using the third dataset. Network models fitted in this research include models for link accidents and for junction accidents. In all cases, coefficients of explanatory variables are studied and the influence of including spatial effects in models are examined. Comparisons of different forms of models are made based on a number of statistical measures, the analysis of residuals and appropriate forms of maps that visualize problems rising from some models and the progress of modelling.

Advantages of models developed in this research compared with conventional models are demonstrated in Chapters 6 and 7. Chapter 8 aims to suggest how these models can be used in practice. It uses two examples to show the possibilities. The first example explains how the models at the local authority level developed in Chapter 6 can be used to predict numbers of accidents in the future. Predictions of accident counts in 2006 based on a conventional model and a CAR model are compared with the observed accident counts obtained from the STATS19 data in 2006. Another example gives details of how to rank sites based on the spatial models for road accidents on a road network. Links on the motorway for which spatial models are developed in Chapter 7 are ranked. High-risk sites selected by the spatial models and the conventional approaches are compared, and implications from the result using different selection criteria are discussed.

The last chapter summarizes the conclusions from this research. Findings and their contributions to the methodology in road safety research are discussed. Limitations of

this research are explained and possible ways to extend this research in the future are suggested.

Chapter 2

Statistical models for road accident data

Much research has been done in the past in order to understand the causes of traffic accidents and to improve road safety. Investigation at the scene can provide detailed information of an individual accident (see Ogden, 1996, Chapter 6). However, an incident such as traffic accident is a random event, so an individual accident may be just a special case. For this reason traffic engineers or policy makers may be more interested in understanding the relationship between accident frequencies and factors such as traffic levels and road geometry, and predicting the total number of accidents in particular areas or on particular roads. These aims can be achieved by using appropriate statistical methods. A number of statistical models have been proposed in the literature. The choice of the analysis approach is mostly determined by the research question and the availability of the data for the accident and other relevant factors.

In this chapter, the general aims of road safety research and the importance of using a statistical modelling approach are discussed in Section 2.1. How accident data are collected and recorded in the UK and how the data can be made to be suitable for statistical models are introduced in Section 2.2. Section 2.3 reviews conventional statistical models for accident frequencies and limitations of these models are discussed in Section 2.4. Previous studies on spatial analysis of road accidents are reviewed in Section 2.5. Section 2.6 introduces the Bayesian modelling approach that is used in this study. How this approach can be applied to develop accident models is briefly explained in Section 2.7.

2.1 The role of statistical models in road safety research

No matter how much is known about the possible generating mechanisms of road accidents, to predict exactly where, when, and to whom the next accident will occur seems to be not practical. However, the total number of accidents during a period, within an area, and of a particular kind may behave with a relatively constant frequency in the long run. Therefore, accident frequencies in the future can be predicted and the relationship between the accident frequencies and some contributory factors can be studied by using appropriate statistical methods.

In order to investigate the causes of road accidents and achieve safer roads by effective means, researchers in road safety may be interested in one or more of the following problems:

- analysing the characteristics of road accidents, including the examination of the association among different characteristics of accidents (for instance, Barker et al., 1998 and Tunaru, 2001);
- identifying spatial clusters of accidents (for instance, Maher and Mountain, 1988 and Flahaut et al., 2003);
- investigating the association between accident frequency, traffic and road geometry (for instance Miaou, 1994, Milton and Mannering, 1998 and Taylor et al., 2002);
- predicting the number of accidents (for instance, Maher and Summersgill, 1996, Mountain et al., 1996 and Greibe, 2003);
- ranking the sites in order of priority for engineering treatment (for instance, Hauer et al., 2004 and Miaou and Song, 2005);
- evaluating safety programmes and engineering treatments (for instance, Wright et al., 1988 and Hirst et al., 2005)

Previous studies in road safety using a statistical approach mainly fall into these categories. In terms of the unit of analysis, there is a difference between the first two types of

studies and the others. To identify spatial clusters and examine the association between different accident characteristics, the unit of analysis is an individual accident. For other types of studies mentioned above, the unit of analysis is usually an area or a location. The statistical modelling approach is the most popular approach applied in such studies. In this, the response variable is normally the total number of accidents in an area or at a location during a fixed period. Therefore accidents need to be aggregated over space and time. How the aggregation can be done will be introduced in the next section.

2.2 General description of the STATS19 data

Many national governments have a department to operate and maintain a national database for road accident data. In order to make statistical analyses of accident data, it is important to understand how the accident data is collected and organised in the database. For instance, in Great Britain, the main source of accident data is the national road accident collection system known as STATS19 (Department for Transport, 2005). Local police forces are responsible for collecting STATS19 data and, in some cases jointly with local authorities, for validating and reporting data to the Department for Transport (DfT). Only accidents involving personal injuries are reported. The STATS19 data consist of three subsets of data, including data of every reported accident, data of every vehicle involved in the accidents and data of every injured individual involved in the accidents. The three datasets are linked to each other via some key variables. The STATS19 data provide information for each recorded accident at a very detailed level and can be used to achieve different research objectives in road safety by appropriate statistical methods.

As explained earlier, when using a statistical modelling approach to analyse accident data, accident data usually need to be aggregated over space and time. The temporal information of each recorded accident consists of year, month, date, day of week and time of day. The spatial information of each accident includes local authority code, location by 10-digit grid references, 1st road number and for junction accidents 2nd road number. After the aggregation of accident data at an appropriate level, the comparison of accident frequencies in different scales of geographical units can be made and also the variation in

the accident data with month of year, day of week, or time of day can be studied.

There are some variables in STATS19 that describe the attributes of the road section on which the accident occurs, for instance number of carriageways, speed limit and type of junction. These variables are needed for accident aggregation when the research interest is to investigate accident frequencies on different road segments, such as junctions and road links. In STATS19, a junction accident is defined as an accident that occurred within 20 metres of a junction. If an accident is coded as a junction accident, the type of the junction at which the accident falls into one of the following main categories: roundabout, crossroads, T- or staggered junction, and multiple junction. Crossroads are defined in STATS19 as 'four arm junctions where the alignments of both roads are uninterrupted whatever the angle of the crossing, and the arms are not staggered'. What are categorised as T-junctions also include '3 arm junctions at which 2 roads join at an acute angle (previously known as 'Y' junction)'. Staggered junctions are 'junctions where several roads meet a main road at a slight distance apart so that they do not all come together at the same point'. Multiple junctions are 'junctions with more than 4 arms (except roundabouts)'.

Another important variable in STATS19 is the severity variable that has three levels of severity—fatal, serious and slight. It is determined by the severity of the most severely injured casualty. Sometimes, accidents aggregated over space and time need to be disaggregated according to different severities. The extent of association between accident frequencies and exposure variables, such as traffic and population, may be different for accidents of different severities. For instance, Jarrett et al. (1989) developed models for accidents of different severities at the local authority level. Their result shows that the estimated coefficients for the explanatory variables can be different when the response variables are for fatal accidents, fatal or serious accidents and accidents of all severity respectively. Therefore, separate models for accidents with different levels of severity are often preferred.

2.3 Statistical models for accident frequencies

Many statistical models have been developed to establish the relationships between accident frequencies, the road environment, traffic levels, and other relevant explanatory variables. For instance, research undertaken by the UK Transport Research Laboratory (TRL) investigated accidents at different types of junction or link in order to determine how accidents are related to vehicle and pedestrian flows, and to the layout and features of junction and road link. Models have been developed for a variety of junctions such as four-arm roundabouts (Maycock and Hall, 1984), rural T-junctions (Pickering et al., 1986), four-arm single carriageway urban junctions with traffic signals (Hall, 1986), three-arm single carriageway urban junctions with traffic signals (Taylor et al., 1996), four-arm priority junctions (Layfield et al., 1996) and three-arm priority junctions on urban single-carriageway roads (Summersgill et al., 1996). Link models have been developed for different types of roads such as urban single-carriageway roads (Summersgill and Layfield, 1996), rural single-carriageway roads and dual-carriageway trunk roads (Walmsley, Summersgill and Binch, 1998; Walmsley, Summersgill and Payne, 1998). Using some of these studies as examples, Maher and Summersgill (1996) gave a comprehensive methodology for the development of predictive accident models. Some technical problems in modelling numbers of accidents were discussed in their paper and solutions to tackle the problems were explained.

When modelling numbers of road accidents, in order to choose an appropriate form of model, the characteristics of accident data should be considered. Generally speaking, the response variable y in a model is the number of accidents at a site, for example a junction, or in a geographical area, for example a county, over a fixed period of time. This indicates that several things need to be considered in order to choose an appropriate form of model. First, the response variable y is an integer and always non-negative. Second, road accidents are random and fairly rare events. Therefore, appropriate forms of probability distribution should be included in models to account for this. Moreover, some factors, such as road-user behaviour and other unmeasured or possibly unrecognized factors, cannot be quantitatively measured, but are believed to affect road safety. Thus, using

appropriate forms of probability distribution can help to take account of the unexplained variation caused by unmeasurable or unrecognised factors.

2.3.1 Poisson model

The Poisson distribution is well-known to describe discrete variables that represent the counts of random events. Therefore, a Poisson model is a natural choice to model numbers of road accidents. Suppose that the number of road accidents at a site (or in a geographical area) in a given period is y, which is Poisson distributed with mean λ:

$$y \sim \text{Pois}(\lambda) \quad (\lambda > 0), \tag{2.1}$$

where λ is the expected number of road accidents at the site in the given period and it varies from site to site. Then, the probability function of y is given by

$$p(y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad (y \geq 0).$$

However, such a model without any explanatory variable does not have any explanatory power. Factors like traffic levels and road geometry contribute to road accidents. Therefore, it is straightforward to extend the model in 2.1 to a Poisson regression model by introducing some relevant explanatory variables. This can be achieved by using the method of generalized linear models (McCullagh and Nelder, 1989). Assume that the numbers of accidents y at different sites are independently Poisson distributed with means λ that depend on features of the sites. For simplicity, suppose that there is only one explanatory variable x, which measures the characteristics of the site. Then the model for numbers of road accidents at a site in a given period can be formulated as:

$$y \sim \text{Pois}(\lambda), \quad \text{where}$$
$$\log \lambda = \beta_0 + \beta_1 x. \tag{2.2}$$

This is a *Poisson log-linear model*, which is a special case of a generalized linear model. It

has been widely used in the statistical literature and has been found to be flexible in fitting different types of count data (e.g. McCullagh and Nelder, 1989). Additional explanatory variables can be included in equation (2.2) and usually are traffic levels, road lengths, etc. It should be noted that for accident models a multiplicative structure of λ has been broadly favoured in the literature (Jarrett et al., 1989), therefore the explanatory variables to be included are usually in logarithmic forms.

Many previous studies in road safety research, especially those in the 1980s, applied Poisson log-linear models. For instance, Maycock and Hall (1984), Pickering et al. (1986), and Hall (1986) studied road accidents on different types of junctions; Jovanis and Chang (1986) examined the relationship between vehicle accidents and vehicle miles of travel; Saccomanno and Buyco (1988) related vehicle accident rates with different traffic volumes, truck types and other relevant factors; Jarrett et al. (1989) compared accident rates between local authorities. There are also some more recent studies applying this type of model (for instance Berhanu, 2004).

One limitation of using the Poisson model is that it assumes that $\text{var}(y) = \text{E}(y) = \lambda$. In other words, the variance of y should equal the expected value λ. But in many applications, count data have been found to display extra variation or *over-dispersion*. That is, the residual variance from the fitted model is often larger than the fitted value. Some possible sources of the over-dispersion in road accident studies were discussed by Miaou and Lum (1993):

- some variables that may have influences on the occurrences of accidents are not included in the model;
- road environment and traffic conditions may not be homogeneous on each road section during a sample period.

Maher and Summersgill (1996) have similarly commented on the occurrence of, and reasons for over-dispersion. If there is an over-dispersion problem, the estimates of the model parameters will still be consistent. In other words, they will converge to the true values when the sample size increases. However, their variances will tend to be under-estimated (see McCullagh and Nelder, 1989).

There are a number of ways in which a Poisson model may be modified to take account of over-dispersion. One way to relax the very restricted rule of equal mean and variance is to introduce an additional parameter τ, which makes $\text{var}(y) = \tau \text{E}(y)$, in the model. This is known as a *quasi-Poisson* model (see Wedderburn, 1974). As discussed by Maher and Summersgill (1996), the parameter estimates resulting from the quasi-Poisson model are identical to those from the Poisson model, but their standard errors are inflated by a factor of $\sqrt{\tau}$. There are some limitations of using a quasi-Poisson model. For instance, the variance is restricted to be proportional to the mean; it needs to be estimated by the method of quasi-likelihood.

2.3.2 Negative binomial (NB) model

Another standard way of modelling over-dispersion is to introduce another level of random variation in λ, modelled by an appropriate probability distribution. Suppose that the value of the true mean λ varies amongst the population of sites according to a gamma distribution. The probability density function of the gamma distribution is defined by

$$f(\lambda) = \frac{\nu^{\kappa}}{\Gamma(\kappa)} \lambda^{\kappa-1} e^{-\nu\lambda},$$

where $\kappa > 0$ is the shape parameter of gamma distribution and $\nu > 0$ is the inverse scale parameter. The mean and variance of λ are κ/ν and κ/ν^2.

Under the assumptions that y has a Poisson distribution with mean λ, while λ follows a gamma distribution $f(\lambda)$, the variability of y over all sites in the population is described by the probability distribution obtained by integrating out λ:

$$q(y) = \int_0^{\infty} p(y|\lambda) f(\lambda) d\lambda. \tag{2.3}$$

This results in a *negative binomial distribution* with parameters κ and ν (see Gelman

et al., 2004, section 17.2), with probability density function

$$q(y) = \frac{\Gamma(\kappa+y)}{y!\Gamma(\kappa)}\left(\frac{1}{1+\nu}\right)\left(\frac{\nu}{1+\nu}\right)^{k}. \tag{2.4}$$

Its mean is $\mathrm{E}(y) = \kappa/\nu$ and variance is

$$\begin{aligned}\mathrm{var}(y) &= \frac{\kappa}{\nu} + \frac{\kappa}{\nu^2} \\ &= \frac{\kappa}{\nu}\left(1+\frac{1}{\nu}\right).\end{aligned} \tag{2.5}$$

The negative binomial distribution allows the mean and variance to be fitted separately. Since both κ and ν are positive, the variance of the negative binomial distribution is always larger than the mean. Therefore, the negative binomial model provides an approach for modelling over-dispersion.

The above mentioned negative binomial model can be extended to a negative binomial regression model in which the expected number of accidents λ_i at a site depends on its characteristics. In the Poisson log-linear model (2.2), an explanatory variable is linked to the Poisson mean via a log-linear model. Remember that more explanatory variables can be included. An extra level of random variation can be introduced in the Poisson mean by including a gamma random effect as expressed in the following equation:

$$\log\lambda_i = \beta_0 + \beta_1 x_i + \log\eta_i, \tag{2.6}$$

where $\eta_i \sim \mathrm{Gamma}(\kappa,\nu)$. Therefore, $\lambda_i = \eta_i\theta_i$, where $\theta_i = e^{\beta_0+\beta_1 x_i}$. If the Poisson mean only depends on θ_i that captures the influence of the explanatory variable, we need the mean of η_i to be 1. Since $\eta_i \sim \mathrm{Gamma}(\kappa,\nu)$, we have $\kappa/\nu = 1$. In this, κ is fixed and does not vary with i. Therefore, according to equation (2.5), $\mathrm{var}(y_i) = \theta_i + \theta_i^2/\kappa$.

A more general kind of NB regression model might be obtained by assuming that $y_i \sim \mathrm{NB}(\kappa_i,\nu_i)$, where κ_i or ν_i might depend on x_i (see Joseph, 2007). Taking κ_i to be a constant but allowing ν_i to depend on x_i leads to a model equivalent to that derived from the multiplicative gamma term (see model (2.6)) with $\mathrm{var}(y) = \theta_i + \theta_i^2/\kappa$. In this,

the variance of y is a quadratic function of the mean. This is the usual form of NB model in applications to road accident data, and belongs to the family of generalized linear models. If ν_i is a constant and κ_i depends on x_i, $\text{var}(y) = \tau\theta_i$, where $\tau = 1 + 1/\nu$. In other words, there is a linear relationship between the variance and the mean. Therefore if ν_i is fixed, the NB model is equivalent to the quasi-Poisson model introduced earlier though the former one is estimated by the method of maximum likelihood and the latter one is estimated by the method of quasi-likelihood.

With the advantage of overcoming the over-dispersion problem in the Poisson models, the negative binomial models have been used in many previous studies in road safety research. For instance, Maher and Summersgill (1996) and Milton and Mannering (1998) used them to model the effects of various highway geometric and traffic characteristics on annual accident frequency on sections of principal arterials in Washington State; Abdel-Aty and Radwan (2000) used them to relate accident occurrence on a principal arterial in Florida with traffic and road geometric characteristics; Berhanu (2004) used them to relate numbers of accidents with road environment and traffic flows on 54 road links in Addis Ababa.

2.3.3 Empirical Bayes methods

The negative binomial model is also known as the *Poisson-gamma model*. Parameters of a negative binomial distribution κ and ν can be estimated using the method of maximum likelihood by fitting the negative binomial distribution to the observed accident distribution for a group of sites. The estimate of the expected number of accidents, λ_i, for site i with a given y_i, can be obtained using the *empirical Bayes method*. Suppose that $\lambda_i \sim \text{Gamma}(\kappa, \nu)$ over all sites and no explanatory variable is taken account of. Using Bayes' Theorem, the posterior density of λ_i can be derived:

$$\begin{aligned} f(\lambda_i|y_i) &\propto p(y_i|\lambda_i) f(\lambda_i) \\ &\propto (e^{-\lambda_i}\lambda_i^{y_i})(\lambda_i^{\kappa-1}e^{-\nu\lambda_i}) \\ &= \lambda_i^{y_i+\kappa-1}e^{-(\nu+1)\lambda_i}. \end{aligned}$$

This is proportional to the Gamma(κ', ν') density, where $\kappa' = \kappa + y_i$ and $\nu' = \nu + 1$.

This means by using a gamma distribution on λ_i, the posterior distribution of λ_i will follow another gamma distribution. Therefore, the gamma distribution is the conjugate family for the Poisson likelihood (see Gelman et al., 2004, Chapter 2). The posterior distribution $f(\lambda_i|y_i)$ reflects the uncertainty about the value of λ_i after taking into account the fact that y_i accidents have occurred at the site. It can be regarded as describing the variation in λ_i amongst those sites at which y_i accidents occurred.

It follows that the mean of the new gamma distribution $f(\lambda_i|y_i)$ is

$$\mathrm{E}(\lambda_i|y_i) = \frac{\kappa'}{\nu'} = \frac{\kappa + y_i}{\nu + 1}.$$

This can be expressed as

$$\mathrm{E}(\lambda_i|y_i) = \left(\frac{1}{1+\nu}\right) y_i + \left(\frac{\nu}{1+\nu}\right) \frac{\kappa}{\nu}.$$

This is known as the *empirical Bayes estimate* of λ. It is a weighted average of the observed count y_i in the site and the overall expected accident frequency κ/ν for all the similar sites in the population. By using the empirical Bayes method, the local mean of the accident count shrinks towards the global mean.

The empirical Bayes estimate introduced above is based on a negative binomial model without any explanatory variable. In a negative binomial regression model, as described in (2.6) in the previous section, $\lambda_i = \eta_i \theta_i$, where $\eta_i \sim \text{Gamma}(\kappa, \kappa)$ and $\theta_i = e^{\beta_0 + \beta_1 x_i}$ (more explanatory variables can be included). In this case, $\lambda_i \sim \text{Gamma}(\kappa, \kappa/\theta_i)$. Therefore, the empirical Bayes estimate of λ_i based on a negative binomial regression model can be written as

$$\mathrm{E}(\lambda_i|y_i) = \frac{\kappa + y_i}{\kappa/\theta_i + 1}.$$

The empirical Bayes method has been used in studies that evaluate the engineering treatment at high-risk sites (Abbess et al., 1981; Hauer et al., 2002; Jarrett et al., 1982;

Wright et al., 1988). This method takes account of the regression-to-mean effect. It improves the conventional approach to comparing accident frequencies at a site before and after a treatment, which can be misleading. There are random fluctuations in accident frequencies. Even if a treatment at a site is not successful, the number of accidents after the treatment might still fall because of the regression-to-mean effect.

2.4 Limitations of conventional models

Similar forms of Poisson log-linear model and the negative binomial model introduced in the previous section have been used by many researchers. However, these models do not take account of the possibility that the response variables might be correlated. Suppose y_{ikt} is the number of road accidents of type k $(k = 1,2,...,K)$, for instance determined by severity, at site i $(i = 1,2,...,N)$ in year t $(t = 1,2,...,T)$. Then, y_{i1t}, y_{i2t},..., y_{iKt} could be correlated. In other words numbers of accidents of different types (for instance, accidents with different severities) at the same site and in the same year could be correlated. This type of correlation exists because y_{i1t}, y_{i2t},..., y_{iKt} will share some common contributory factors. If any of these factors is unmeasured or unobserved, the correlation will not be fully explained by the model. On the other hand, some unmeasured contributory factors might be almost constant over time, therefore numbers of road accidents of type k at site i y_{ik1}, y_{ik2},...,y_{ikT} could be correlated. This is known as *temporal autocorrelation*. When both the temporal correlation and the correlation between different types of accident exist, more complicated correlation would be introduced, for instance, there should be a correlation between the numbers of accidents of different types in different period at site i, y_{i11} and y_{i22}. Moreover numbers of road accidents at different sites, namely i and j, but of the same type k and in the same year t, which are y_{ikt} and y_{jkt}, could be correlated. This is known as *spatial autocorrelation*. Some contributory factors tend to have similar levels of values at neighbouring sites. Therefore, they are often spatial correlated. If any of these factors is unmeasured or unobserved, the spatial correlation in the response variables will not be fully explained by the model. The spatial correlation is more complicated to take account of than the temporal correlation. This is because the temporal

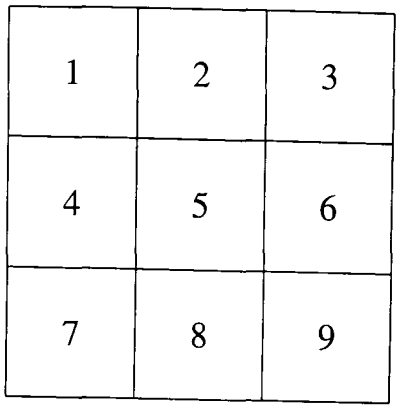

Figure 2.1: Rectangular grids for areal models.

correlation is one-dimensional while the spatial correlation is usually two-dimensional. Compared with the other two types of correlation, spatial autocorrelation has not been considered much in accident models in the past. However, it needs to be taken account of in both areal models and network models for accidents. Further details of why it exists and needs for consideration are explained below.

For areal models, the shapes of local authorities or wards are usually irregular. However, in order to illustrate the spatial information implied by the geographical maps, rectangular grids are used here to represent the locations and the connectivity of the areas. Suppose the accident counts y_i $(i = 1, ..., 9)$ during a fixed period at 9 areas are known. These areas correspond to the 9 numbered grids in Figure 2.1. y_i is often assumed to be Poisson distributed with a mean λ_i. Conventional accident models treat the accident means λ_i in different areas as independent and ignore the spatial information in the maps. However, areas especially those that share common boundaries are not spatially independent. The extent of development and urbanization for adjacent areas is more likely to be similar. In addition, in the context of road accidents, traffic moves on the roads, accidents occur on the roads and adjacent areas always have some common roads passing between them. Therefore, the assumption of independent accident means is not appropriate for areal models. A node-link graph has been plotted over the rectangular grids in Figure 2.2. It illustrates the dependence relationship in the 9 grids. Grids whose centroids are connected via a direct link are treated as neighbours. The accident means λ_i for neighbouring grids is assumed to be correlated. The correlation between different grids can be

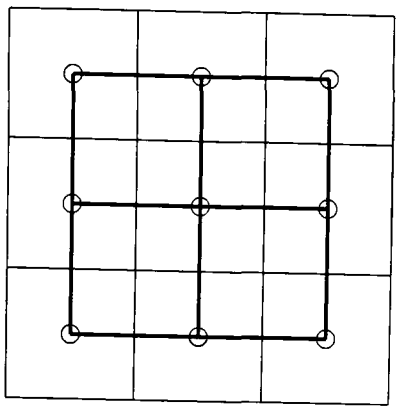

Figure 2.2: Spatial dependency among the grids.

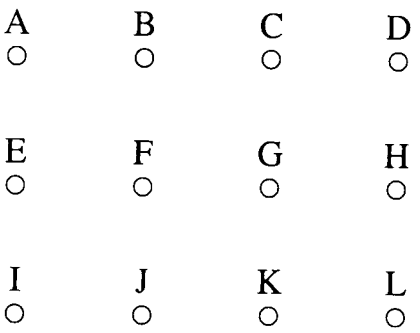

Figure 2.3: A node graph for junctions.

expressed as $\text{cor}(\lambda_i, \lambda_j) = \rho_{ij}$, and in a simple case we might assume $\rho_{ij} > 0$ if i and j are neighbours and is 0 otherwise. Moreover, when modelling numbers of road accidents on a road network, the spatial structure of the road network and the spatial dependence within the road network were seldom considered in previous studies. Suppose the nodes in Figure 2.3 represent a certain type of junctions on a road network. Conventional models for junction accidents treat the accident means at different junctions as independent (so $\text{cor}(\lambda_i, \lambda_j) = 0$) and ignore the spatial information implied by the road network. However, numbers of road accidents at neighbouring sites in a road network are more likely to be correlated than those at non-contiguous sites, especially when neighbouring sites have some common physical and environmental features, such as similar road geometry and traffic flow. In other words, previous research in studying road accidents at the road network level does not take full account of the structure of the existing road network, and therefore, may lose a chance to find further insight of the effect of road network design and urban planning on the occurrence of road accidents. When the spatial structure of the road network is considered, the spatial dependence relationship in the sites can be identi-

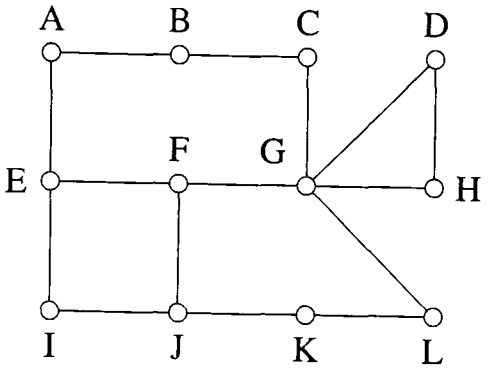

Figure 2.4: Spatial dependency among the nodes.

fied. In Figure 2.4, any link connecting the nodes represents there is a direct road between the junctions. Nodes that are connected directly via a link are defined as neighbours and are spatially correlated. Such a link-node graph illustrates the spatial dependency among the junctions. This spatial information can be used to introduce the spatial correlation in the expected numbers of accidents for the neighbouring sites, expressed as $\rho(\lambda_i, \lambda_j)$. Similarly, when developing models for link accidents, account should also be taken of the spatial dependency among neighbouring links.

Failure to take account of the temporal autocorrelation and the spatial autocorrelation may cause some problems for model estimation. Schabenberger and Gotway (see 2005, section 1.5) have a general discussion on the effects of autocorrelation on statistical inference. First, when autocorrelation in the data is ignored, the evidence against the null hypothesis that the coefficient is zero can be overstated—the test rejects more often than it should. This is because the variability of the coefficient is under-estimated. After taking account of the correlation in a linear model, the estimator can be more variable (less precise) than the estimator for independent data. Secondly, the effect of positive autocorrelation is that n correlated samples provide less information than n independent samples.

Normally, if any type of the correlation introduced above exists while the model does not take it into consideration, some problems can be identified by checking the residuals from the model. For instance, Jarrett et al. (1989) developed statistical models for accidents in 4 years. A study of residuals from their models showed that there was significant correlation both in time and in space. When spatial or temporal correlation is found in

the residuals from the models, it will indicate that the original model needs to be improved by considering the correlation (that is not explained by the explanatory variables) in the response variables. This implies a multivariate model is needed. However, there is no natural Poisson multivariate model to achieve this. A straightforward approach is to assume that $y_{ikt}|\lambda_{ikt}$ (the number of road accidents y_{ikt} given the mean λ_{ikt}) are independently Poisson distributed but that the λ_{ikt} are correlated. If the Poisson log-linear model 2.2 is used, then introducing some appropriate forms of correlation in the second level of the model is straightforward. This can be done by including some random effects in the model. The structure of such random effects should reflect the appropriate dependent relationship in the λ_{ikt}. For instance, to model the spatial autocorrelation, spatially correlated random effects should be included. Using the conventional frequentist approach to fit such models with random effects is very difficult, especially when more than one type of correlation present and the structure of the random effects is very complicated. There are some recent studies that use a Bayesian modelling approach, which is explained in a later section, and take account of different types of correlation discussed above. For instance, Tunaru (2002) developed multivariate models for accidents of different types; Miaou et al. (2003) included both spatial effects and temporal effects in their models.

2.5 Spatial analysis of road accidents

Accident data are spatial data, and include locational information such as a geographical grid reference. This type of information enables the spatial analysis of accident data. Previous studies that consider the spatial aspect of the accident data include the identification of locations where accidents clustered and developing models that take account of spatial correlation in neighbouring areas like local authorities or neighbouring sites like junctions or links.

2.5.1 Spatial cluster identification

Some work has been done in spatial analysis of road accidents using methods of studying spatial point processes. These aim to investigate whether the distribution of road accidents follows some systematic process so as to form a clustered or regular pattern and therefore identify areas or road segments where accidents cluster. This kind of areas and road segments are also known as blackzones, high-risk sites or hotspots.

A variety of methods for point pattern analysis have been used to study accident data. To test the existence of general clustering in a point pattern, the nearest neighbour distances and the K-function are two usual approaches (see Schabenberger and Gotway, 2005, section 3.3). The approach of nearest neighbour distances compares the observed nearest neighbour distance with the distance under complete spatial randomisation. Levine et al. (1995*a*) used it to study the spatial patterns of motor vehicle accidents in Honolulu. The K-function measures how many events (for instance, accidents) occur within a certain distance of other events. Jones et al. (1996) used K-function to determine the degree of clustering exhibited by the residuals from a spatially referenced logit model constructed to ascertain the factors influencing the likelihood of death in a road traffic accident. Their study aims to test if there was some systematic geographical factor influencing the outcome not adequately controlled for in the model. More recently, Okabe and Yamada (2001) proposed a network K-function method that extends the ordinary K-function to a network space, where locations of events are restricted to a network and distances are measured as a network distance. Yamada and Thill (2002) compared the ordinary and the network K-function in traffic accident analysis. Their result indicates that the planar K-function tends to over-detect clustering patterns which are random patterns in the sense of network K-function.

To investigate the presence for and locations of local clusters, a kernel density function (see Silverman, 1997) can be used. It requires a band width that determines the size of the kernel and the overall smoothness of the resulting estimate. Flahaut et al. (2003) studied locations of blackzones in a particular road in Belgium based on the kernel density function approach. Anderson (2007) used the kernel density estimation to create a density

surface for visualizing the hotspots in London. Like the planar K-function, this approach ignores the network structure and might over-detect clusters.

2.5.2 Models with spatial aspects

One common limitation of the studies introduced above is their constraint to a descriptive analysis instead of a statistical modelling approach. Although the statistical approach to modelling the accident counts has been available for a long time, conventional models seldom take account of any spatial aspects that are based on the structure of the road network or the contiguity of the geographical areas. However, as explained in Section 2.4, there are some reasons to believe that the counts will not be statistically independent especially at neighbouring sites.

Maher (1987) suggested that spatial autocorrelation between the mean accident frequencies at neighbouring sites may account for the apparent migration of road accidents from treated sites to untreated sites, as observed by Boyle and Wright (1984). Loveday and Jarrett (1991, 1992) attempted to measure the amount of spatial autocorrelation in some real data sets for road networks in different regions of England. Moran's I (Upton and Fingleton, 1985), which measures the amount of spatial autocorrelation between data on a mapped network, was used to estimate the correlation between the accident frequencies (and hence between the λs) at neighbouring sites on a road network. The extent of spatial correlation in the accident data was very different for road networks in different regions. The studies suggested that the magnitude of the autocorrelation appeared to depend on the complexity and density of the road network.

These studies indicate that there might be spatial correlation in the observed accident counts at sites on a road network. Therefore, the mean of the accident frequencies at neighbouring sites might be correlated. Loveday and Jarrett (1992) gave some possible reasons for the existence of the spatial correlation. Firstly, the level of traffic flowing through neighbouring sites will not vary markedly. Accident frequencies are related to traffic flow, so one would expect mean accident frequencies at neighbouring sites to be related. This means that $\mathrm{cor}(\lambda_i, \lambda_j) = \rho_{ij}$ could be high when sites i and j are close.

Secondly, the physical characteristics of the road environment itself have a degree of similarity between neighbouring sites, and so the accident potential tends to be less varied than it is between distant sites on the network. This implies that there is less variability in the mean accident frequencies between neighbouring sites than between sites chosen at random from the network. Including explanatory variables, such as traffic levels and road geometry, in the models may explain some of the spatial correlation in the accident data. But it is still possible that there are other unmeasured or unobserved features of the sites that are related to accident frequencies left out of the models. If these unmeasured features are similar for the neighbouring sites, the spatial correlation in the accident data may not be fully explained by such models. This may result in spatially correlated residuals. Similar problems can occur for models of accidents at the area level.

Loveday and Jarrett (1992) illustrated a way to take account of the spatial correlation in a linear network by moving average models. Consider the linear network shown in Figure 2.5. The network represents a simple road network with each node representing a junction and each link representing a road link. Suppose again the observed count of accidents at a junction or on a link is Poisson distributed with mean λ_i. u_i and v_i are the random effects associated with each node and each link respectively. If these random effects are gamma distributed, the mean λ_i for node i can be modelled as

$$\lambda_i = u_i + u_{i-1} + u_{i+1} + v_{i-1} + v_i. \tag{2.7}$$

This model implies that the expected number of accidents at node i is the sum of the random effect at the node itself and the random effects at the adjacent nodes and links. It is a discrete form of moving average model (see Best et al., 2005). With the same idea, a model for accidents on a link can also be developed. More random effects should be included in the moving average model 2.7 when the linear network in Figure 2.5 is extended to a grid network. The moving average models suggested by Loveday and Jarrett (1992) have some limitations. First, no explanatory variable was considered. Secondly, a road network is more complicated than the regular networks assumed in their study.

Levine et al. (1995*b*) developed a spatial lag model at the census block level to ex-

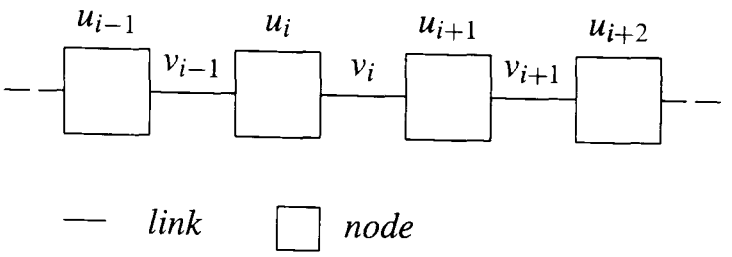

Figure 2.5: A linear network.

amine the relationship of motor vehicle accidents to population, employment and road characteristics. In their model, the response variable y_i is determined not only by the explanatory variables but also by the other y_j, where site j is site i's neighbour. However their model is based on an assumption that the accident count is normally distributed, which is a serious limitation.

Miaou et al. (2003) developed spatial models for traffic crashes in Texas at the county level. Later, they extended their study to multivariate models (Song et al., 2006) which model different types of accidents jointly. The models in both of their studies consider the spatial dependency in the counties. Their approach to taking account of the spatial dependency is known as the conditional autoregressive (CAR) model. It is also the most important technical concept in this thesis. Its definition, why it is appropriate to be applied to models for road accidents and how it can be added to the conventional models to take account of the spatial correlation in the Poisson accident means will be introduced in the next chapter. Although the two studies introduced above successfully use the CAR model to develop models for the accident count, they have a few limitations. First, the studies only include a limited number of explanatory variables, which are actually surrogate variables. Second, Texas counties have more regular boundaries than is the case in England, which simplifies some of the spatial modelling. Last, their papers do not show any result of residual analysis for models with and without spatial effects.

2.6 Bayesian methods for data analysis

As shown in the end of Section 2.4, multivariate models for accidents of different type and models with spatial or temporal effects are complicated and difficult to formulate and fit to data in the context of the frequentist approach. Fortunately, the Bayesian approach provides an effective solution and comprehensive framework to solve this problem. Two principal approaches to inference that guide the modern data analyst are the frequentist approach, and the Bayesian approach. 19th century science was broadly Bayesian in its statistical methodology, while frequentism dominated 20th Century scientific practice (Bradley, 2005). The frequentist evaluates procedures based on repeated sampling, imagining an infinite replication of the same inferential problem and evaluating properties over this repeated sampling framework for fixed values of unknown parameters. The Bayesian requires a sampling model and, in addition, a prior distribution on parameters. Unknown parameters are considered random and all inferences are based on their distribution conditional on observed data, which is known as the *posterior distribution*.

The Bayesian approach permits the use of previous knowledge or subjective opinion in specifying a prior distribution. Therefore, frequentists often criticise Bayesian procedures for their loss of objectivity by using specific priors. *Noninformative priors* with densities that can be described as vague, flat or diffuse, are some Bayesians' response. Bayesian statistics has seen a strong movement away from subjectivity towards objective uninformative priors in the past 20 years (Bradley, 2005).

For complicated models, the Bayesian approach facilitates the inclusion of several random effects by formulating them in different layers via a hierarchical structure. Recent progress in Markov Chain Monte Carlo method and the computer implementation of it make computation of Bayesian models more convenient and faster.

2.6.1 Applying Bayes' Theorem

In the Bayesian approach, in addition to specifying the model for the observed data $y = (y_1, y_2, ..., y_n)$ given a vector of unknown parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$, θ is supposed to be a random quantity having a prior distribution $\pi(\theta|\eta)$, where η is a vector

of hyperparameters. Inference concerning θ is then based on its posterior distribution $p(\theta|y,\eta)$, given by

$$p(\theta|y,\eta) = \frac{p(y,\theta|\eta)}{p(y|\eta)} = \frac{p(y,\theta|\eta)}{\int p(y,\theta|\eta)d\theta} = \frac{p(y|\theta)\pi(\theta|\eta)}{\int p(y|\theta)\pi(\theta|\eta)d\theta.} \tag{2.8}$$

This formula is referred as *Bayes' Theorem* (Carlin and Louis, 1996). The result of the integral in the denominator is actually the marginal distribution of the data y given the value of the hyperparameter η. If η is unknown, the proper Bayesian solution would be to quantify this uncertainty in a second stage prior distribution. Denoting this prior distribution by $h(\eta)$, following Carlin and Louis (1996), the posterior for θ is now obtained by marginalizing over η,

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{\int p(y,\theta,\eta)d\eta}{\int\int p(y,\theta,\eta)d\eta d\theta} = \frac{\int p(y|\theta)\pi(\theta|\eta)h(\eta)d\eta}{\int\int p(y|\theta)\pi(\theta|\eta)h(\eta)d\eta d\theta.} \tag{2.9}$$

Alternatively, η can be replaced by an estimate $\hat{\eta}$ obtained as the value which maximizes the marginal density of $p(y|\eta)$. Inference is now based on the estimated posterior distribution $p(\theta|y,\hat{\eta})$. Such an approach is often referred to as empirical Bayesian analysis. Section 2.3.3 has explained how to get empirical Bayes estimate of the mean frequency of road accidents at a site.

The hyperprior for η can also depend on a collection of unknown parameters resulting in a generalization of expression (2.9) featuring another stage of prior. This procedure of specifying a model over several levels is called *hierarchical modelling*, with each new distribution forming a new level in the hierarchy. However, when models get complicated via a hierarchical structure, difficulties in computation rise at the same time.

2.6.2 Bayesian computation

Implementation of the Bayesian approach requires the calculation of posterior distributions. The calculation often involves complex and high-dimensional integrals like those in expression (2.9), in which θ and η are usually vectors. The most popular computing tools in modern Bayesian analysis are *Markov Chain Monte Carlo* (MCMC) methods. MCMC

methods make Bayesian computation relatively easy by reducing the high-dimensional integration problem to a series of lower-dimensional problems. MCMC is essentially *Monte Carlo integration* using *Markov chains*. The original Monte Carlo integration was developed by physicists to use random number generation to compute integrals (Gilks et al., 1996). Its characteristics make it popular in Bayesian analysis to approximate posterior distributions. Monte Carlo integration draws samples from the required distribution, and then forms sample averages to approximate expectations. With a similar idea, MCMC methods approximate posterior distributions based on samples generated by running one or more Markov chains. The idea of Markov Chain simulation is to simulate a process, which will converge to a stationary distribution after a long enough period of time.

There are many ways of constructing the required Markov chains, but the most popular algorithms are the *Gibbs sampler* and the *Metropolis-Hastings algorithm*. Suppose a model features k parameters, $\theta = (\theta_1, ..., \theta_k)'$. The Gibbs sampler requires that samples can be generated from each of the full conditional distributions $p(\theta_i|\theta_{j\neq i}, y), i = 1, ..., k$. Under mild conditions, the collection of full conditional distributions uniquely determine the joint posterior distribution $p(\theta|y)$ (see Banerjee et al., 2004, section 4.3.1). The Gibbs sampler is easy to implement, but sometimes one or more of the full conditional distributions may not be available in a closed form. If this is the case, the Metropolis-Hastings algorithm can provide a solution. Now samples will be drawn from a joint posterior distribution $p(\theta|y)$ rather than any full conditional distribution. The algorithm requires a rejection step from a particular candidate density at each iteration. Based on a ratio of probabilities (see Banerjee et al., 2004, section 4.3.2), either the new simulated value is used for the current state or the value in the previous state remains.

For both of the algorithms, enough time is needed until convergence to the target stationary distribution of parameters can be achieved. This time period is often known as *burn-in period*. Samples in the burn-in period will not be used to produce the estimates of parameters. Convergence diagnosis is a very important step in Bayesian analysis using MCMC methods. However, it is difficult to decide when it is safe to stop the simulation and summarize the output. If the simulated Markov chain has not converged to the sta-

tionary distribution, the inference can be wrong. The most common approach is to run several parallel chains simultaneously, starting from different initial points (Spiegelhalter et al., 2003). These chains are then plotted together against the axis that represents number of iterations. These trace plots are usually used to monitor the status of convergence of simulated chains by observing the extent of overlapping in these chains. But there are some problems with this approach. For instance, the process of monitoring requires a subjective judgment by the observer. Some existing convergence diagnostic statistics can provide more formal approaches for convergence diagnosis. If convergence is attained then the empirical distribution of each chain should be almost identical to the empirical distribution obtained by pooling all the chains together. If convergence is not reached, the variations within each chain are smaller than the variation within the pooled chain. Based on this idea, Gelman and Rubin (1992) developed the *Gelman-Rubin statistic*, which calculates a ratio regarding to the between-chain variance and within-chain variance. When the simulated Markov chain converges to the stationary distribution, the Gelman-Rubin statistic should approach 1. Later, this statistic is modified by Brooks and Gelman (1998) by calculating a ratio of lengths of between-chain interval and within-chain intervals, which is considerably simpler than the original method. Calculation of this ratio, namely $\hat{R}$, is implemented in both WinBUGS (Spiegelhalter et al., 2003) and the R2WinBUGS (Sturtz et al., 2005) package for R. WinBUGS is the Windows version of BUGS (Bayesian inference Using Gibbs Sampling) for fitting Bayesian models.

2.6.3 Estimation of parameters and model comparison

As suggested by Carlin and Louis (1996), for a typical Bayesian data analysis, we might summarize our findings on estimated parameters by reporting (1) the posterior mean, (2) several important posterior percentiles (for instance, at the levels 0.025, 0.50, and 0.975), (3) a plot of the posterior itself if it is multimodal, highly skewed, or otherwise badly behaved, and possibly, (4) posterior probabilities of the form $P(\theta > c|y)$, where θ represents a parameter and c is some important reference point that arises from the context of the problem.

As in the frequentist approach, a plot of predicted values against observed values can help to illustrate the fit of a model. A good fit would show the points evenly scattered around the line with a 45 degrees slope. Examining the posterior distributions for parameters of interest and hyperparameters are also very helpful for model checking. For instance, suppose y is the response variable. In order to check the fit of a model to data, we could draw simulated values from the posterior distribution of y^* (y^* is the predicted value of y using the model) and compare these samples to the observed data (for example, to examine the probability that the samples contain the observed data). This is known as *posterior predictive check* (see Gelman et al., 2004, Chapter 6).

Moreover, examination of residual helps to identify any problems with the model, for instance, temporal or spatial correlation. However, the approach to calculating residuals is more complicated in a Bayesian context. Suppose, a linear model with response variable y is fitted using N observations. In a frequentist context, residuals are $y - y^*$, where y^* are fitted values, and have N values. If a Bayesian approach is used, for each observation, y^* will have a posterior distribution depending on values saved from each simulation. Therefore, if M simulations are saved for model estimation, we will have a $N \times M$ matrix for y^* therefore will obtain a $N \times M$ residual matrix. Albert and Chib (1996) name the elements of this matrix *Bayesian residuals* and suggest using these to examine the fit of binary response regression models. In Chapter 4, it is shown how models can be checked by examining the extent of spatial correlation in these residuals.

In addition to the above mentioned methods for model checking, some formal approach is needed to measure the performance of the candidate models and choose those that perform better. There exist some well-known criteria for Bayesian model comparison. A natural approach is to compare models based on the posterior probability of the model given the data. This can be achieved by calculating the ratio of the observed marginal densities for two candidate models. The ratio is known as the *Bayes factor* and it indicates which model is favoured by the data (Banerjee et al., 2004, section 4.2). However, the Bayes factor is often difficult to calculate and not suitable for high dimensional models. Akaike (1974) introduced a criterion for models comparison in a frequentist framework,

namely *Akaike Information Criterion* (AIC). The formula is

$$AIC = -2\log(\text{maximized likelihood}) + 2p,$$

where p is the number of estimated parameters. The model with the smallest AIC is preferred. An alternative to AIC is the *Bayesian Information Criterion* (BIC) (Schwarz, 1978). BIC adjusts AIC by including the sample size n in the calculation as

$$BIC = -2\log(\text{maximized likelihood}) + p\log n.$$

Both of AIC and BIC are penalized model choice criteria, since they penalize a model by its complexity measured by the number of parameters in the model. However, under the condition of using any noninformative prior, none of these model comparison criteria may be appropriate (Banerjee et al., 2004, section 4.2).

Spiegelhalter et al. (2002) proposed the use of the *Deviance Information Criterion* (DIC), which is a natural generalization of AIC. The generalization is based on the posterior distribution of the *deviance* $D(\theta)$ (McCullagh and Nelder, 1989), which is defined as -2 times the log-likelihood ($\log p(y|\theta)$) plus a constant, where θ are parameters of the models. The posterior expectation of the deviance, denoted by $\bar{D}$, is calculated using $\bar{D} = \int D(\theta)p(\theta|y)d\theta$. The effective number of parameters P_D is a measure of the complexity of the model and is defined by

$$\begin{aligned} P_D &= \bar{D} - D(E_{\theta|y}[\theta]) \\ &= \bar{D} - D(\bar{\theta}). \end{aligned}$$

In this, $\bar{D}$ equals the sample mean of the simulated values of $D(\theta)$ and $D(\bar{\theta})$ is the deviance calculated by replacing θ with its posterior expectation $\bar{\theta}$. The DIC is defined as $DIC = \bar{D} + P_D$, which combines a measure of fit together with the effective number of parameters. Therefore, like AIC and BIC, DIC is a criterion based on a trade-off between the fit of the data to the model and the complexity of the model. The model with the

smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed. This suggests that any reduction in DIC is desirable, but it is difficult to judge what would constitute an important difference in DIC. Spiegelhalter et al. (2003) suggest that 'differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC'. The advantage of DIC over other criteria, for Bayesian model selection, is that the DIC is easily calculated from the samples generated by a Markov chain Monte Carlo (MCMC) simulation.

2.7 Bayesian models for numbers of road accidents

Consider again the Poisson log-linear model in 2.2, within the Bayesian framework. The model can be completed in the form of

$$\begin{aligned} y_i &\sim \text{Pois}(\lambda_i), \\ \log \lambda_i &= \beta_0 + \beta_1 x_i, \\ \beta_j &\sim \text{independent } \text{N}(0, \sigma^2), j = 0, 1. \end{aligned} \tag{2.10}$$

This formulation using a sequence of parameters and priors constitutes a Bayesian hierarchical model. Under a fully Bayesian framework prior distributions for the parameters have to be set up. For β_0 and β_1, a normal prior of mean 0 and variance σ^2 is applied. Normally, σ^2 is set to be very large like 10000 to ensure the priors on the parameters are noninformative. More explanatory variables can be included in the model with independent normal priors on their coefficients.

Additional random effects ε_i can be included in the Poisson mean λ_i. If ε_i is the log of a gamma-distributed variable as shown in equation (2.6), then the marginal distribution for y_i is negative binomial. Therefore, the model is the same as a negative binomial regression model introduced in Section 2.3.2. It is a Bayesian version of the negative binomial regression model, with normal priors on the regression coefficients. However, this model

is difficult to be extended to a more complicated form, for instance multiple response models that jointly model different types of accidents. In such circumstances, a log-normal random effect is usually preferred. In this, ε_i is normally distributed. The marginal distribution of y_i is then Poisson-lognormal that is analytically intractable therefore cannot be solved using maximum likelihood method. This is not a problem in a Bayesian context by using Markov Chain Monte Carlo methods (see Section 2.6.2).

When ε_i is normally distributed, it is often formulated as

$$\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$
$$\sigma_\varepsilon^2 \sim h(),$$

where $h()$ is an appropriate hyper-prior distribution for σ_ε^2. Various noninformative prior distributions have been suggested for $h()$ in the Bayesian literature, including an improper uniform prior and an inverse-gamma prior. Gelman (2006) explores and makes recommendations for prior distributions for variance parameters in the hierarchical model. He also suggests that care should be taken when the inverse gamma prior is used since the resulting inferences will be sensitive to the choice of the parameters in the gamma prior.

ε_i is an unstructured random effect. Some structured random effects can also be included in the model. Tunaru (2002) developed multivariate models that allow for correlations between mean accident frequencies for different severity. The correlation was introduced in the variance-covariance matrix for the random effect. Miaou et al. (2003) employed hierarchical Bayesian models similar to those in disease mapping to build risk maps for area-base traffic crashes in Texas at the county level. Their models considered spatial correlation in the counties by introducing spatially structured random effects. More recently, Bailey and Hewson (2004) applied a multivariate modelling method to model a range of traffic safety performance indicators simultaneously. Some approaches to taking account of spatial and other types of random effects in Bayesian models will be introduced in the next chapter.

Chapter 3

Bayesian models for spatial data

Limitations of conventional models for accident frequencies have been discussed in the previous chapter. These models assume independence between different sites or areas. Spatial correlation either in the road network or between adjacent geographical areas is not well considered in these models. Accident data are spatial data. Researchers in areas such as ecology and epidemiology have shown particular interest in applications with spatial data. A variety of spatial models have been developed in these areas. Some of these models can be applied to road accident data, but the models may need some modification in order to make them more appropriate for accident data.

This chapter aims to explore some existing approaches to the development of statistical models for spatial data and to explain which of these approaches are appropriate to model accident frequencies. How univariate spatial models can be extended to multivariate models and to include temporal effects will be discussed. Moran's I statistic and the influence of ignoring edge effects in spatial models will be introduced at the end of chapter.

In order to find an appropriate approach to account for spatial effects in the conventional models, recall the Poisson log-linear model with log-normal random effects

introduced in Section 2.7:

$$y_i \sim \text{Pois}(\lambda_i), \quad \text{where}$$
$$\log \lambda_i = \beta_0 + \beta x_i + \varepsilon_i.$$

The ε_i are unstructured random effects and independent in space. Spatial correlation can be introduced by replacing ε_i with spatially structured random effects that can express spatial dependence in all the sites.

Previous studies that develop accident models based on this idea are very few. This is because the recent development of spatial modelling approaches in other areas have not been widely used in road safety research. Therefore, in the following section, some existing statistical modelling approaches for spatial data will be introduced and the possibility of generalizing these approaches to model road accident frequencies will be discussed.

3.1 Bayesian spatial models

3.1.1 General introduction of spatial data analysis

In the previous chapter, the general approach for Bayesian modelling has been explained. Some applications in Bayesian analysis involve data with additional spatial information, for instance, geographically referenced data. Such data are known as *spatial data*. Researchers in areas such as ecology and epidemiology have shown particular interest in this type of data. Based on applications with different purposes, Banerjee et al. (2004) classify spatial data sets by three basic types, namely *point-reference data*, *areal data* and *point pattern data*.

In the case of point-referenced data, the response variable y is a vector that measures values of some factor at some locations in a study region. The value of this factor should vary continuously by location. One example is to model the amount of some pollutant in the air monitored in a number of sites in an area. When data for n locations are available,

a general model can be:

$$y|\mu, \theta \sim \mathrm{N}_n(\mu, \Sigma),$$

where N_n denotes the n-dimensional normal distribution, μ is the mean level and Σ is the variance-covariance matrix for y. Σ describes the spatial correlation in different locations. The simplest approach to identifying Σ is to make it depend on the distance between locations. Banerjee et al. (2004) give a variety of choices to produce distance-based variance-covariance matrix. Many studies in this area aim to predict some values at unobserved locations based on observed data in several known locations.

Banerjee et al. (2004) call the second type of spatial data areal data. When the whole area of interest can be partitioned into a finite number of areal units (of regular or irregular shape) with well-defined boundaries, models can be developed at the geographical area level. Many studies in disease mapping belong to this category of application. Best et al. (2005) compared a class of Bayesian spatial models for disease mapping. In their study, models are grouped in three categories, namely models with correlated normal random effects, semi-parametric spatial models and spatial moving average models. The first type of model includes models with joint normal priors and models with conditional autoregressive (CAR) priors. Many applications in areas such as disease mapping use such models. The second type of model assumes the whole study region can be partitioned into k clusters of areas. There are some methods to choose cluster locations and allocate areas to clusters. For instance, k areas are randomly chosen as cluster centres and the remaining areas are allocated to a cluster if this cluster centre is closer than others in terms of the minimal number of area boundaries that have to be crossed to reach it. k is treated as unknown and the relative disease risk in each cluster is constant. The relative risk for each cluster is then modelled by a gamma distribution or a log-normal distribution. The last type of models are spatial moving average models. They have been developed primarily for continuous processes. Best et al. (2000) proposed a discrete version of the gamma moving average model and applied it to model a kind of illness. In such models, it needs to specify a number of grid cells that define the latent process in order to model

the spatial dependence structure.

The last type of Banerjee et al. (2004)'s classification is spatial point pattern data. Here, the response variable y often represents the occurrence of an event. The location of the event is assumed to be random. The main interest with data of this kind is to identify locations where some specific events cluster.

The three types of applications using spatial data, introduced above, cover a wide range of spatial data analysis. It is worth thinking about how accident data fit into this classification. The first type of application based on point-referenced data may be not appropriate for accident data, since it needs the response variable to vary continuously in the space. However, the last two types of applications are relevant for accident data. Models for accidents at the area level, such as local authority and ward, use a kind of areal data. Methods to analyse point pattern data can be used to identify clusters of road accidents. Some relevant studies in this area have been reviewed in Section 2.5.1.

Since this study aims to develop accident models that take account of spatial effects, the remaining part of this section will emphasise the introduction of two spatial modelling approaches that are used in disease mapping, namely the conditional autoregression model and the spatial moving average model.

3.1.2 Conditional autoregressive (CAR) model

Research in disease mapping often uses areal data. Generally, disease mapping aims to explain the geographical distribution of disease rates, and to identify areas with low or high rates. Bayesian methods are currently much applied in this area. Ghosh and Rao (1995) conducted a comprehensive review of hierarchical Bayesian methods and found them favourable for dealing with small area estimation problems when compared with other statistical methods. The conventional approach does not take account of any spatial dependency in disease. In other words, the response variables in different areas are treated as independent. In disease mapping (e.g. Mollié, 1996), the response variable y_i is often the number of deaths or specific disease cases. It is assumed to be Poisson distributed with a mean $E_i r_i$. In this, E_i is the expected number of deaths or cases based on the age-sex

distribution of area i and standard rates for the event or condition. r_i is the relative risk for area i, linked to one or more explanatory variables via a log transformation. When only one explanatory variable x_i is available, the model can be expressed as $\log(r_i) = \beta_1 x_i + \varepsilon_i$, where ε_i is supposed to be independently and normally distributed with mean zero and constant variance σ^2. This model is very similar to the Poisson model with a log-normal random effect for the accident count that has been introduced in Section 2.7 in the previous chapter. The variance-covariance matrix of the ε has the form $V = \sigma^2 I$. The independence assumption will be violated when the error terms are autocorrelated. This problem frequently happens for spatially located data as well as for data arranged in time sequence. For instance, in disease mapping, geographically close areas may tend to have similar disease rates. The variance-covariance matrix will then have nonzero off-diagonal elements reflecting dependence between the outcomes of neighbouring areas. There are two approaches to modelling the dependence between neighbouring areas. One is the *simultaneous autoregressive model* (SAR), in which spatial outcomes are expressed in terms of a joint density function for area i and its neighbours together. Another approach is to express the spatial structure via a *conditional autoregressive model* (CAR). In the absence of any explanatory variable for each area, suppose $\log(r_i) = \theta_i$. Then, by the CAR specification, the conditional distribution of each θ_i given all the other $\theta_j (j \neq i)$ depends only on its neighbours. Banerjee et al. (2004) give some reasons why using the CAR model to formulate spatial dependency structure shows more advantages than using the SAR model in the Bayesian context. The CAR model is computationally convenient by Gibbs sampling, which is, like the CAR model, based on full conditional distributions. Moreover, in practice a SAR specification is not used in conjunction with a generalized linear model. In disease mapping, a Bayesian estimate of the disease rate in an area, based on a CAR model, not only shrinks to the global mean as an empirical Bayes estimate introduced in Section 2.3.3 but also shrinks towards a local mean, according to the rates in the neighbouring areas.

The *intrinsic Gaussian autoregression* model is one way of formulating a CAR model for irregular maps. In this context, the conditional mean and variance are defined in terms

of a square matrix of non-negative weights $W = \{w_{ij}\}$ that describe the association in the observed areas. The simplest choice for W is $w_{ij} = 1$ if areas i and j are neighbours, and $w_{ij} = 0$ otherwise. The conditional distribution of each θ_i is then given by

$$\theta_i|\theta_j\ [j \neq i] \sim \mathrm{N}\left(\sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}}\theta_j, \frac{\tau_\theta}{w_{i+}}\right), \tag{3.1}$$

where $w_{i+} = \sum_{j=1}^{N} w_{ij}$ and τ_θ is a scale parameter; $j \in N[i]$ means j is a neighbour of i. τ_θ / w_{i+} is the conditional variance. It is inversely proportional to w_{i+}, that is the total number of neighbours for area i if the 1–0 weighting scheme is used. Although in this formulation the conditional distributions $\theta_i|\theta_j, j \neq i$ are proper, the corresponding joint distribution $p(\theta)$ is improper, whose integral is infinite. This indicates that this intrinsic Gaussian model can be used only as a prior in a Bayesian analysis. However, the impropriety of the joint distribution can be remedied by introducing an extra parameter ρ, which will make the conditional distribution of θ_i become

$$\mathrm{N}\left(\rho \sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}}\theta_j, \frac{\tau_\theta}{w_{i+}}\right). \tag{3.2}$$

But is the proper CAR formulation better than the improper CAR? The answer may be no. Banerjee et al. (2004) give some of the reasons. Firstly, in a proper CAR model, the mean of θ_i is not an average of its neighbours, but some proportion of this average. Does this make any sensible spatial interpretation? Moreover, proper CAR models can make the range of spatial pattern restricted while improper CAR models may enable a wider scope for posterior spatial pattern. Therefore, the choice between two types of CAR models is ambiguous. However, care should be taken when including improper CAR priors because it can make the posterior joint distribution improper, so that the resulting posterior distribution makes Bayesian inference impossible. The propriety of the posterior joint distribution, when improper priors are included in the models, has been studied by many researchers. For instance, Ghosh et al. (1998) as well as Sun et al. (1999) provided a sufficient condition to gain a proper joint posterior with a univariate CAR prior for spatial random effects. It needs that the response variable y is always positive. This will not be

a problem for areal models of accidents as long as the area is large enough. However, for network models, especially when only one year's data are used, number of accidents at a junction or on a link could be zero. Therefore, further research is needed for examining the propriety of the posterior joint distribution when the response variable contains zero.

The formulation of θ_i only based on a CAR model implies a high degree of spatial interdependence. It may be modified to allow for a mixed or compromise scheme where some variation is explained by an unstructured term, which describes unstructured heterogeneity in the relative risks. This can be expressed as:

$$\log(r_i) = \theta_i + \varepsilon_i$$

$$\theta_i|\theta_j\ [j \neq i] \sim \mathrm{N}\left(\sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}}\theta_j, \frac{\tau_\theta}{w_{i+}}\right)$$

$$\varepsilon_i \sim \mathrm{N}(0, \tau_\varepsilon),$$

where τ_θ and τ_ε are the variance parameters for the spatially structured random component θ and unstructured random component ε. Such a model, incorporating both the spatial random effects and the unstructured random effects on the log relative risks, is called a *convolution Gaussian model* (Besag and Mollié, 1989; Besag et al., 1991). As described by Mollié (1996), τ_θ and τ_ε control the strength of θ_i and ε_i respectively. If $\tau_\varepsilon/\tau_\theta$ is close to $\bar{w}$, the average value of w_{i+} in equation (3.1), v_i and θ_i have the same importance. If $\tau_\varepsilon/\tau_\theta$ is larger than $\bar{w}$, then unstructured heterogeneity dominates; if it is smaller, spatial structured variation dominates. Depending on the relative strength of unstructured as against spatially structured variation, individual area risks will be smoothed towards the global or neighbourhood averages.

The above mentioned existing method used in disease mapping is straightforward and very suitable to apply for modelling of road accidents at the geographical area level, for instance local authority or ward.

3.1.3 Spatial moving average models

The CAR model is an extension of the Poisson log-linear model by including spatially structured random effects. Best et al. (2000) proposed a form of Poisson regression model with identity link instead of the logarithmic link. Their approach allows for incorporating spatial dependence via a spatial moving average model for the latent random effects. This is similar with the moving average models proposed by Loveday and Jarrett (1992) to model accidents on a road network as introduced in Section 2.5.2. Best et al. (2000) used a Poisson-gamma spatial moving average model to analyse traffic-related air pollution and respiratory illness in children living in the Huddersfield region of northern England. The study region was partitioned in two ways, namely census enumeration districts and regular grid cells. The general models in their study are:

$$y_i \sim \text{Pois}(\lambda_i)$$

$$\lambda_i = N_i r_i$$

$$r_i = \beta_0 + \beta_1 x_i + \beta_2 \sum_j k_{ij}\gamma_j. \tag{3.3}$$

For each area i, y_i is the number of cases of self-reported frequent cough amongst children aged 7-9, N_i is the estimated population of 7-9 year old children and x_i is the average annual nitrogen dioxide concentration. The γ_j are gamma random variables, where $\gamma_j \sim$ Gamma(κ_j, ν), that can be thought of as latent unobserved risk factors associated with sub-regions or locations indexed by j. These sub-regions or locations are typically defined by the user. In this example, they can be either the same areal partition as the disease outcome data or an arbitrary partition of the area using grids. The k_{ij} are elements of the kernel matrix and measure the relative contribution of the latent variables γ_j to the random effect in area i. Best et al. (2000) assume a stationary Gaussian kernel function with $k_{ij} = 1/2\pi\phi^2 \exp(-d_{ij}^2/2\phi^2)$, where d_{ij} represents the Euclidean distance between the centroid of area i of the study region and the location of the jth latent factor and ϕ is the spatial range parameter governing how rapidly the influence of the latent gamma random variables on the area specific excess risk declines with distance. Independent

gamma prior distributions are assumed for the parameters β_0, β_1, β_2 and γ_j as this enables the MCMC sampler to exploit conjugacy with the Poisson likelihood. The specification of these priors has been discussed in detail by Best et al. (2000).

Although the spatial moving average model is relatively easy to apply to spatial data, it has some limitations to use in practice. Normally, latent grids are chosen to represent a partition of the region that is appropriate for capturing unmeasured spatial variation in the disease rate. But there is no general rule to guide how to define the latent grid. Best et al. (2000) suggested that, for the Poisson-gamma spatial moving average model, the output areas of the study region and the grids for the latent process need not be regular grid cells. However, according to the concept of the moving average model, this type of model is more appropriate to be used for a region that is partitioned into regular grids. Accidents are seldom studied at the grid level. This is possibly because the data for the explanatory variables are often not available at the grid level. Such data are normally collected for each areal unit that is defined by a particular type of boundary, such as local authority and ward. Moreover, analyses based on grid cells imply the arbitrary partition of the road network. Therefore, areal models for road accidents are most likely to be developed at the area level like local authority and ward.

3.2 Extensions of univariate CAR models

3.2.1 Multivariate CAR models

The CAR model introduced in the previous section models accidents of one particular type. It can be extended to model accidents of different types jointly. There are several possible ways to extend univariate CAR models to multivariate CAR models. An obvious choice would be to use several separate univariate CAR models, each of which with a common scale parameter τ for the spatial random effect θ_i. If the variation in θ_i is different for accidents of different types, different τ's should be applied. However, the response variables, for instance, numbers of fatal and serious accidents, could be correlated since they may share some same risk factors. Therefore, a multivariate CAR model, which

takes account of not only the spatial dependence between neighbouring areas but also the correlation in the mean frequencies for different types of accident, is more appropriate. It is expected to improve the estimate of the mean number of a particular type of accident at location i by not only borrowing information from neighbouring locations but also using information from other types of accidents.

Tunaru (2002) developed Bayesian models for multiple count data. These models take account of the correlation in the Poisson means of different types. Suppose that the unstructured random effects in area i are $\varepsilon_{1i},...,\varepsilon_{pi}$, where p describes the accident type. They are distributed with a multivariate normal distribution N_p, whose variance-covariance matrix takes account of the relationship in different types of accident and is modelled via a Wishart distribution. It is straightforward to extend Tunaru's models to multivariate CAR models by including additional random components that capture spatial correlation in the response variables. Several formal multivariate CAR models have been proposed in the literature. Kim et al. (2001) proposed a "twofold CAR" model for modelling two types of disease over each areal unit. However, their method is limited to bivariate data and is difficult to generalize to a large number of diseases. Based on Mardia (1988)'s fundamental theory for multivariate Gaussian Markov random fields (GMRF), Gelfand and Vounatsou (2003) and Carlin and Banerjee (2003) suggested a class of multivariate proper CAR models. Jin et al. (2005) proposed generalized multivariate conditional autoregressive (GMCAR) models for areal data, in which the joint distribution for the multivariate spatial process is defined through simple conditional and marginal forms.

Generally, multivariate CAR models can be formulated using a similar idea to the univariate CAR model. Suppose the spatial effect θ_i in area i now is a p-dimensional vector, which corresponds to p types of incidents. Analogous to the univariate CAR case, the full conditional distributions of θ_i given θ_j, where $j \neq i$, are $N_p(\sum_{j \in N[i} \frac{w_{ij}}{w_{i+}}\theta_j, \frac{1}{w_{i+}}\Sigma)$, which is a multivariate Gaussian distribution with p dimensions. The $p \times p$ matrix Σ is positive definite and represents the conditional variance-covariance matrix with the pth diagonal element representing the conditional variance of the pth component of θ_i and

off-diagonal elements representing the conditional covariances between each pair of the p elements of θ_i.

3.2.2 Models with spatio-temporal effects

In disease mapping, when longitudinal data are available, models can include both spatial effects and temporal effects. Most of the Bayesian methods extend the CAR model of Besag et al. (1991) to a spatio-temporal model. Bernardinelli et al. (1995) suggest a model of the following form:

$$\log(r_{it}) = \alpha + \delta_i t + \theta_i + \varepsilon_i.$$

α is the intercept (overall rate). θ has an intrinsic Gaussian auto-regressive prior with variance σ_θ^2. ε_i is normally distributed with mean 0 and variance σ_ε^2. $\delta_i t$ models a linear trend t with the coefficient δ_i depending on i. δ_i can either be taken as an overall average growth rate δ or modelled via an intrinsic Gaussian prior with variance σ_δ^2 if the trend is expected to have a spatial structure. The latter choice introduces spatio-temporal interactions in the model.

Waller et al. (1997) use a model like that of Besag et al. (1991), but apply it to each time point separately. The model allows that the scale parameter of the spatial component varies in different years.

Knorr-Held and Besag (1998) adopt the following model,

$$\log(r_{it}) = \alpha_t + \beta_t x_{it} + \theta_{it},$$

where α_t follows a random walk and $\theta_{it} = \theta_i$ for all $t = 1, \ldots, T$. They use the same spatial random effect in all the years for each observational unit. Their model also allows coefficients of the explanatory variables to vary in time. However, the model combines temporal and spatial effects additively and does not allow for space and time interactions.

Most of the recent studies in disease mapping develop their spatio-temporal models based on one of the approaches introduced above. More recently, Knorr-Held (2000)

introduces four types of space and time interaction in his models. The number of deaths y_{it} in county i during year t is assumed to have a binomial distribution with parameters n_{it} and π_{it}. n_{it} is the number of persons at risk and π_{it} is modelled with a logit link as $\eta_{it} = ln(\pi_{it}/(1-\pi_{it}))$. η_{it} decomposes additively into time- and space- dependent effects as

$$\eta_{it} = \mu + \alpha_t + \gamma_t + \theta_i + \varepsilon_i + \phi_{it},$$

where α_t is a temporal effect modelled by a random walk, γ_t is an unstructured temporal effect, θ_i is a spatial effect modelled by a CAR prior, ε_i is an unstructured spatial effect and ϕ_{it} describes the spatio-temporal interaction. Four types of interaction were considered by Knorr-Held (2000) by interacting two temporal effects with two spatial effects respectively. They are the interaction between the unstructured temporal effect γ_t and the unstructured spatial effect ε_i, the interaction between the random walk effect α_t and the unstructured spatial effect ε_i, the interaction between the unstructured temporal effect γ_t and the spatially structured effect θ_i and the interaction between the random walk effect α_t and the spatially structured effect θ_i.

The most up-to-date study for Bayesian multivariate models with spatio-temporal effects is by Richardson et al. (2006). In their study, the joint pattern of the spatio-temporal variation of males and females lung cancer risks in four periods is analysed at the ward level. Their models extend the shared component spatial models by Knorr-Held and Best (2001). Let y_{1it} and y_{2it} represent the observed number of cases of lung cancer in males and females, respectively, for ward i and time period t. As usual, they are assumed to be Poisson distributed with mean $E_{1it}r_{1it}$ and $E_{2it}r_{2it}$. E_{1it} and E_{2it} are the expected number of lung cancer cases calculated on the basis of average age-sex specific incidence rates for the region. Without considering any explanatory variables, Richardson et al. (2006) propose the following models to account for shared temporal and spatial components in males and females, differential temporal and spatial effects in males and females, space-

time interaction terms and heterogeneity terms in the relative risk r_{1it} and r_{2it}:

$$\begin{aligned} r_{1it} &= \theta_i \delta + \xi_t \kappa + \eta_{it} + \phi_{1it} \\ r_{2it} &= \frac{\theta_i}{\delta} + \frac{\xi_t}{\kappa} + \upsilon_i + \psi_t + \eta_{it} + \phi_{2it}. \end{aligned} \tag{3.4}$$

In these models, θ_i represent the shared spatial pattern in males and females, υ_i represent the spatial pattern in females that is different from that in males, ξ_t is a shared time trend, ψ_t is the female differential from the male time trend, η_{it} captures the space-time interaction. Improper CAR priors as formulated in 3.1 are used for the spatial random effects θ_i and υ_i. The first order random walk priors are assumed for the time trends ξ_t and ψ_t. The authors use the CAR priors to formulate these temporal effects by identifying the neighbour structure according to time. The heterogeneity terms $(\phi_{1it}, \phi_{2it})^t$ are assigned a zero-mean multivariate normal distribution to allow for the correlation between the male and female disease process in each space-time unit.

3.3 Moran's I statistic

Moran's I is often used as a measure of the spatial correlation in the data. A standard formulation of Moran's I in data y (see Upton and Fingleton, 1985) is

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of data and w_{ij} is the spatial weight between location i and j. Moreover, a *spatial correlogram* shows changes of spatial correlation in the data when distance or order of neighbours is increased. In general, the spatial correlation should drop with the increase of distance or order of neighbours.

The randomization procedure that is broadly favoured in testing the significance of spatial correlation can be replaced by the normality approach if the data to be examined are residuals from regression models. There is a procedure in the R 'spdep' package (Bivand, 2004) for testing Moran's I when it is calculated from residuals from a linear

regression. But for residuals from a generalized linear model, no ready to use adjusted method can be found for Moran's I. Ge and Zhang (2006) did some simulations to examine the performance of different type of residuals from a Poisson log-linear model in the Moran's I test and suggest the use of Pearson residuals, defined as

$$\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}, \tag{3.5}$$

where $\hat{\lambda}_i$ is the estimate of the Poisson mean using method of maximum likelihood. It is found to perform better than the regression residual $y_i - \hat{\lambda}_i$ in a number of different conditions for generalized linear models.

In a Bayesian context, a simple approach to examining the spatial correlation in residuals could be to calculate Moran's I, by using the posterior mean of λ_i for $\hat{\lambda}_i$ in equation (3.5). However, as mentioned in Section 2.6.3, in a Bayesian context, Bayesian residuals are more appropriate to be used for model checking. Especially in the case of examining spatial correlation in residuals, only residuals obtained from λ_i saved from the same simulation will completely reflect the real spatial pattern in residuals. However, no available approach is found in the literature that examines the spatial correlation in Bayesian residuals. Gelman et al. (2000) proposed a general approach for model checking using posterior predictive simulations. This suggests examining the posterior distribution of Moran's I. In this case, more computations will be needed. A general introduction to Gelman's approach and how it can be applied to Moran's I statistic for model checking are explained in the next chapter.

3.4 Edge effects

In disease mapping, spatial analyses are usually undertaken within a region that is contained within a boundary. Data for the areas outside the boundary are often missing. Therefore, estimates for areas close to the edge can only be based on available data within the study region and could be statistically biased. Vidal Rodeiro and Lawson (2002) ex-

plored how the estimation of the relative risk from a disease at or near boundaries can be affected by the level of the missing data for areas outside the edge by fitting a number of models, including the Poisson-gamma model, Poisson log-normal model, intrinsic CAR model and convolution CAR model, with some simulated data. Vidal Rodeiro and Lawson (2002) also reviewed some effective solutions that have been used to remove or compensate for edge effects. Constructing an external buffer zone is one of the most popular approaches for accommodating edge effects. If observations within the buffer zone are available, it is straightforward to fit models with the whole data covering both the study region and the buffer zone. If no data are observed in the buffer zone, ordinary methods for the missing data problem can be used.

Chapter 4

Methodology: Spatial models for accident frequencies

Conventional statistical approaches for modelling numbers of road accidents have been reviewed in Chapter 2. The Poisson log-linear and negative binomial models are widely used in road safety research. One limitation of these models is the spatially independence assumption. Possible reasons for the existence of the spatial correlation in the accident frequencies have been discussed in Section 2.5.2. However, there are relatively few studies that have considered spatial correlation in their models.

Spatial models have been developed and extensively studied by researchers in areas such as disease mapping. One approach to taking account of spatial dependence is to use the CAR model, which has been introduced in Section 3.1.2. The very similar context for disease and road accidents indicates that this approach can be used for road accident models.

In this chapter, the approach to modelling the number of road accidents is proposed. Models are introduced in order of increasing complexity. The univariate models include the Poisson model with log-normal random effects, the Poisson model with log-normal random effects and fixed regional effects, the Poisson model with log-normal random effects and spatially structured random effects modelled by a CAR prior and the Poisson model with log-normal random effects, spatially structured random effects and temporal

effects. For CAR models, possible choices for the neighbours list and the spatial weighting scheme in the context of accident models are discussed in detail. Later, the univariate models are extended to multivariate models that model numbers of different types of road accidents jointly. Methods for model checking are described at the end.

4.1 Univariate models

A simple form of model for accident data is the Poisson log-linear model. This model can be extended to more complicated models. In the following subsections, univariate models will be proposed. Based on a simple version of the Poisson log-linear model, unstructured random effects, spatial effects and temporal effects will be introduced. Notation used in this chapter is defined below:

- y_i is the total number of a particular type of road accidents at site i in a fixed period;
- β_0 is the intercept;
- x_i is an explanatory variable which measures some characteristics of the site in the same period; β_1 is the coefficient for it. For simplicity, only one explanatory variable is used when models are introduced. But in practice more explanatory variables can be added by extending $\beta_1 x_i$ to $\sum_{j=1}^{p} \beta_j x_{ij}$, where p is the number of explanatory variables.

4.1.1 Poisson log-linear model

A Poisson log-linear model can be formulated as:

$$\begin{aligned} y_i &\sim \text{Pois}(\lambda_i), \\ \log \lambda_i &= \beta_0 + \beta_1 x_i, \end{aligned} \tag{4.1}$$

where y_i are independent over space given λ_i. As introduced in Section 2.3.1, for count data like numbers of accidents, the residual variance is often larger than the mean. However, a Poisson model requires that the variance equals the mean and therefore cannot

take full account of the variation in the accident data. One way to improve the model is to introduce extra variation in the Poisson mean.

4.1.2 Poisson regression model with log-normal random effects

As explained in Section 2.3.2, the negative binomial model can provide a solution for the overdispersion problem by introducing an extra level of random effects, that follow a gamma distribution, in the Poisson mean. However, it is difficult to be extended to a more complicated form of model (see Section 2.7), for instance, multivariate models or models with spatial random effects. A Poisson regression model with log-normal random effects is more flexible in this context. It can be expressed in the following form:

$$\begin{aligned} y_i &\sim \text{independent Pois}(\lambda_i) \\ \log \lambda_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \end{aligned} \tag{4.2}$$

where the ε_i are independently and normally distributed with mean 0 and constant variance σ_ε^2. This model assumes the λ_i over all areas or sites are independent. However, this may not be true. Possible reasons are discussed in Section 2.4. Therefore, this model needs to be modified in order to take account of the spatial dependency among the areas or sites.

4.2 Univariate spatial models

4.2.1 Poisson regression model with regional (fixed) effects

One way to take account of the spatial dependency is to introduce spatial fixed effects in model (4.2). When the data used to fit the model are at the areal level like local authority or ward, the data normally cover contiguous areas that are located in the whole study region. The region could consist of several sub-regions by which the areas can be grouped. For instance, suppose the whole study region is the West Midlands region of England (not including rural areas) and the areal units are wards. Then the sub-regions

are the metropolitan districts Birmingham, Dudley, Sandwell, Solihull, Walsall, Wolverhampton and Coventry. With this information about sub-regions, a factor represented by dummy variables can be created and included in the model to measure the regional effects. Suppose that there are M sub-regions. M dummy variables can be defined as:

$$D_{iL} = \begin{cases} 1 & \text{if } i \in S_L, \text{where } L = 1, 2, \cdots, M \\ 0 & \text{otherwise,} \end{cases}$$

where $S_1, S_2, \cdots, S_M$ represent identified sub-regions. $i \in S_L$ means area i is in S_L. These variables will capture the fixed regional effects. However, only $M-1$ dummy variables will be needed in a model. Based on model (4.2), an extra term $\sum_{L=1}^{M-1} \alpha_L D_{iL}$ can be added in the right hand side of the second line. α_L is the coefficient of the dummy variable D_{iL}.

The fixed regional effects can account for the spatial dependency among the areas on their sub-regions therefore introduce spatial dependency at a relatively wide level. For better explaining the spatial dependency, a method for taking account of the spatial correlation among the areas at a more local level is needed.

4.2.2 Poisson regression model with spatial random effects

The intrinsic Gaussian prior, introduced in Section 3.1.2, can be used at this stage to formulate a model for spatial random effects that takes account of the spatial dependency in the neighbouring areas or sites. It should be noted that in the context of disease mapping, based on the same distribution assumption (Poisson distribution) of the response variable, the log-linear model is for the relative risk r_i for area i, where $y_i \sim \text{Pois}(E_i r_i)$. E_i is the expected number of disease cases based on the age-sex distribution of area i. But when the response variable is the accident count, such a formulation is not appropriate. This is because the age-sex distribution is not a very appropriate indicator of the expected number of road accidents in an area. It might be appropriate for models of road casualties since some particular population groups, for instance school children, might be high-risk. For such models casualties at the area level could be aggregated in two ways in terms of

the casualty's home location or of the accident location. Hewson (2005) used the CAR model to study child pedestrian injuries at the ward level in Devon County. He found evidence that data aggregated in terms of the casualty's home location cannot be assumed to be spatially independent. Conversely, data aggregated in terms of the accident location were found to be spatially independent. The spatial dependence in the casualty home aggregation data can be understood as implying that casualties tend to have accidents near their home, given estimates of the distance between home and accident location of around 600m, but not necessarily within their home ward.

Some researchers like Miaou et al. (2003) use a traffic variable to take place of E_i and treat r_i as the accident rate in models for accident count. However, this approach can be questioned because the approach is based on the assumption that the expected number of accident in an area is proportional to the amount of traffic. This assumption could be too strict. It is more reasonable to include the traffic variable in the log-linear model like other explanatory variables.

The spatial CAR model used in this research is formulated as:

$$
\begin{aligned}
y_i &\sim \text{Pois}(\lambda_i) \\
\log \lambda_i &= \beta_0 + \beta_1 x_i + \theta_i \\
\theta_i | \theta_j \, [j \neq i] &\sim \text{N}\left(\sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}} \theta_j, \frac{\tau_\theta}{w_{i+}} \right),
\end{aligned} \tag{4.3}
$$

where, as explained in Section 3.1.2, $w_{i+} = \sum_{j=1}^{N} w_{ij}$ and τ_θ is a scale parameter; $j \in N[i]$ means j is a neighbour of i. As shown in the previous chapter, the formulation of model (4.3) implies a high degree of spatial interdependence and may be modified to allow for a mixed or compromise scheme where some variation is explained by an unstructured term ε_i, which is normally distributed with mean 0 and constant variance σ_ε^2.

The CAR model defined above is improper so it can only be used as a prior distribution for the spatially distributed random effects. It is often convenient to assume that such random effects have sum to zero mean. Besag and Kooperberg (1995) show that constraining the random effects to sum to zero and specifying a separate intercept term with

a location invariant Uniform$(-\infty, +\infty)$ prior is equivalent to the unconstrained parameterisation with no separate intercept. In WinBUGS, the improper CAR model includes a sum-to-zero constraint on the spatial random effects therefore an intercept with an improper uniform prior is needed. This can be done by assigning the intercept with the dflat() distribution (used in WinBUGS), which corresponds to an improper flat prior on the whole real line. This means that in model (4.3) $\beta_0 \sim$ dflat().

A prior distribution for the overall variance parameter τ_θ needs to be specified. In WinBUGS, the inverse variance parameter, also known as the precision parameter, is used when a CAR model is specified. A prior distribution needs to be assigned to it. A gamma prior is usually used. As suggested by Spiegelhalter et al. (2003), one option is to use a gamma distribution with shape and inverse scale parameters both equal to 0.01. It has a mean of $0.01/0.01 = 1$ and a large variance of $0.01/(0.01)^2 = 100$. This tends to place most of the prior mass away from zero for the standard deviation of the spatial random effects. When the true spatial dependence between areas or sites is negligible (for instance, the standard deviation is close to zero), this may induce artefactual spatial structure in the posterior. Kelsall and Wakefield (1999) suggest an alternative Gamma$(0.5, 0.0005)$ prior for the precision parameter of the spatial random effects in a CAR model. This expresses the prior belief that the random effects standard deviation is centred around 0.05 with a 1% prior probability of being smaller than 0.01 or larger than 2.5. This prior is adopted to specify the prior distribution for the inverse variance parameter in a CAR model in this research.

4.2.3 Spatial neighbours list and weighting choice

4.2.3.1 Neighbours list

Using a conditional autoregressive model to formulate the spatial random effect as in model (4.3) requires an appropriate specification of the spatial neighbours list and the weighting scheme.

For geographical areas, neighbours of area i can be defined as other areas that share at least one common boundary with it. Both GeoBUGS (Thomas et al., 2004), which is an

add-on in WinBUGS, and the 'spdep' package (Bivand, 2004) in R are able to work out the neighbours list based on this definition. In R, a neighbours list is identified by examining regions with contiguous boundaries. Under the 'queen' condition (Upton and Fingleton, 1985), a single shared boundary point meets the contiguity condition. GeoBUGS includes a tolerance zone of 0.1 metres when it examines the contiguity condition. There will be no problem if the units of analysis, such as local authorities, in the study region are completely contiguous. However, if, for instance, a river passes between two local authorities and there is no shared boundary point for the local authorities, they will not be identified as neighbours.

Figure 4.1: Incomplete map of London boroughs

Figure 4.1 shows part of London, including the boundary lines between boroughs and the layout of A-roads, based on the Meridian 2 Ordnance Survey data (Meridian^{TM}, 2007). The grey lines correspond to the boundaries and the black lines plot the layout of the A-roads. Hammersmith and Fulham, Kensington and Chelsea, Westminster, the City of London and Tower Hamlets are on the north side of River Thames while Wandsworth, Lambeth, Southwark and Lewisham are on the south side of the river. The boroughs on

the different sides of the river will not be identified as neighbours in GeoBUGS and the 'spdep' package in R because their boundary lines are separated by the river. Using such a neighbours list implies that the boroughs on one side of the river are independent from the boroughs on the other side of the river. However, this may not be true when the response variables are numbers of road accidents. Some A-roads are found to cross the river from one borough to another. For instance, there are two roads connecting Hammersmith and Fulham with Wandsworth; City of London is connected to Southwark by three roads. Under these circumstances, the underlying means of number of the road accidents in these boroughs may not be independent because some traffic could move between them through the roads that connect them. With some minor changes in the neighbours list produced by GeoBUGS or R, the modified list, based on the connectivity by roads, is more reasonable to be used in a CAR model to describe the spatial dependency.

Neighbours found by a contiguity condition are *first order* neighbours. *Second order* neighbours or even *higher order* neighbours can be included in the neighbours list. Generally, if area j is a first order neighbour of area i, then the first order neighbours of j (excluding those that are also first order neighbours of i) are defined as the second order neighbours of i. Higher order neighbours are defined in the same sense. High order neighbours can be used in a CAR model especially when the unit of analysis is relatively small area like a ward.

There are also other types of neighbour definition. For each area in the study region, the approach of k nearest neighbours finds the closest k areas as its neighbours. The distance based approach finds neighbours within some given distance. Euclidean distance between the centroids of the two areas are often used. However, these two approaches are more suitable for regular maps. Therefore, they are not considered here.

The approaches to identifying a neighbours list discussed above are for geographical areas. For models of accidents on a road network, a neighbours list should be determined by the structure of the road network. Figure 4.2 plots an artificial road network. Major roads are plotted in black and minor roads are plotted in grey. The points correspond to the location of some accidents. A standard approach to identifying a neighbours list for

Figure 4.2: Accidents on a road network.

such a road network is to replace it with a node-link-cell system. The network is normally defined by major roads, and nodes are intersections between the roads, or possibly points on a road where there is a change in the speed limit or number of carriageways. Links are the sections of road between nodes. Accidents on the main roads are then allocated to a node or a link. Accidents allocated to a node are normally junction accidents, which are conventionally defined as those occurring within 20 metres of a junction. Accidents on minor roads are allocated to a cell of the network. Figure 4.3 illustrates the result for the network of Figure 4.2. Numbers in the figure correspond to total numbers of the accidents allocated to a node, a link or a cell. When only junction accidents are considered, a neighbours list needs to be identified for the five nodes, namely A, B, C, D and E. Nodes i and node j are first order neighbours if they are connected by a direct link. For instance, nodes B and E are first order neighbours for node A. When only link accident are considered, two links can be defined as first order neighbours if they join in a same node. For instance, links AB and BC are neighbours because they join in a same node B. When both junction accidents and link accidents are considered, the neighbours

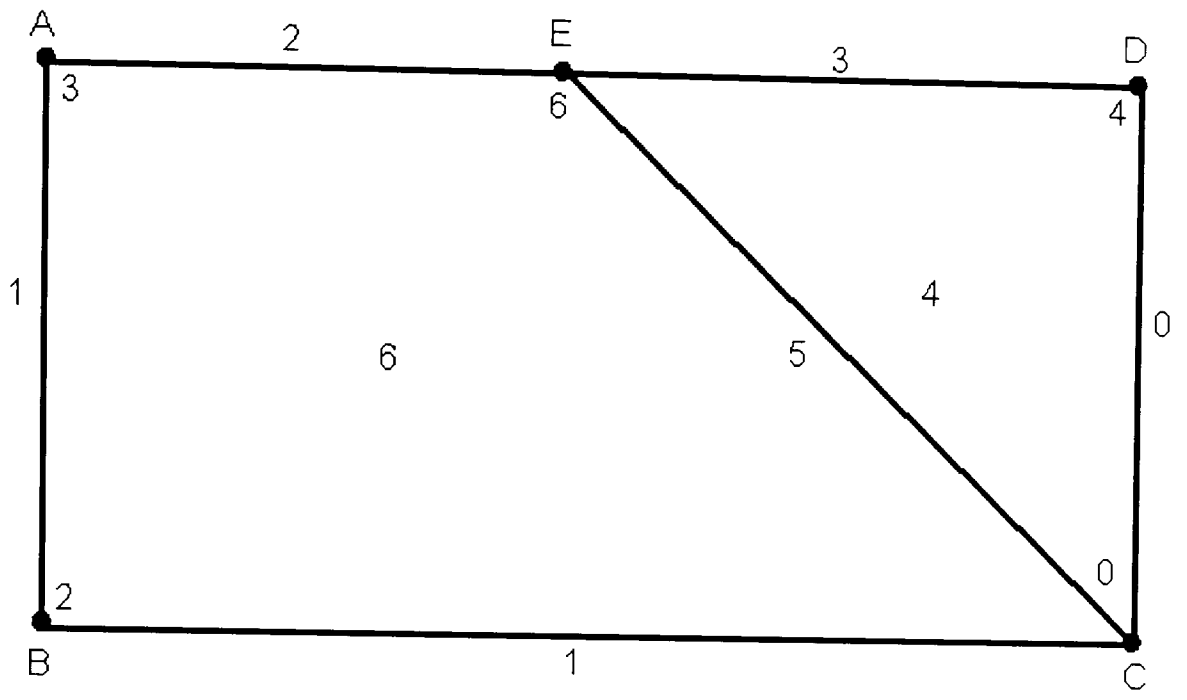

Figure 4.3: A node-link-cell system.

list for the whole road network will be more complicated. Under these circumstances, the first order neighbours for a node include other nodes that are connected by a direct link with this node and the links that join in this node; the first order neighbours for a link include other links that have a common node with this link and the nodes that are on each end of this link. Using this definition of neighbours, node A has 4 first order neighbours, namely nodes B and E and links AB and AE; link AB also has 4 first order neighbours, namely links BC and AE and nodes A and B. Higher order neighbours can also be considered in the neighbours list for a road network.

As illustrated in Figure 4.3, cells are areas bounded by the major roads. Therefore, neigbours for a cell consist of cells that share a common major road with it. However, no previous study is found to analyse accidents at the cell level. This could be due to the reason that explanatory variables at this level are difficult to obtain.

4.2.3.2 Weighting choice

In order to specify a CAR prior, the simplest choice for the weight matrix W is $w_{ij} = 1$ if areas or sites i and j are neighbours, and $w_{ij} = 0$ otherwise. When areal models are considered, an alternative for the weights is to let them depend on the Euclidean distance between the centroids of the neighbouring areas, the smaller the distance, the larger the weight. In network models, the distance based weights are more appropriate for models of junction accidents. In these, the weights depend on the distance, measured by the length of the road link between the neighbouring junctions. For areal models, it is also possible to use the percentage of shared boundary for area i with its neighbours to determine the weight, but this is difficult to work out in GIS software.

In areal models, the above weighting choices do not take account of the spatial structure of the road network and therefore may lose some valuable information to determine the spatial weights. Traffic moves on roads and road accidents occur on the roads. Traffic is the most significant factor to explain the variation in accident frequencies. However, it is very difficult to obtain a perfect measurement of traffic. This implies that even models that include traffic variables can leave some unmeasured quantity due to the difficulty in measuring traffic. Such an unmeasured quantity may be partly explained by a CAR model with a suitable weighting plan. As explained earlier, the spatial layout of roads can be used to complement the neighbours list and make the definition of neighbours more appropriate in the context of modelling the number of road accidents. This information can be also considered to determine the spatial weights that reflect the extent of spatial correlation. The greatest the number of common roads that cross two neighbouring areas, the higher should be the correlation in the traffic volume in the areas. A reasonable explanation for this is, with more common roads passing through the neighbouring areas, there will be more traffic movement between the areas. Therefore, for area i, a higher weight can be given to its neighbour j if there are more common roads passing through i and j. The amount of the weight can be determined by the total number of common roads.

Figure 4.4 is a map of part of Southern England in the 1980s showing both the local authority boundaries and the main road network. Blue lines represent motorways.

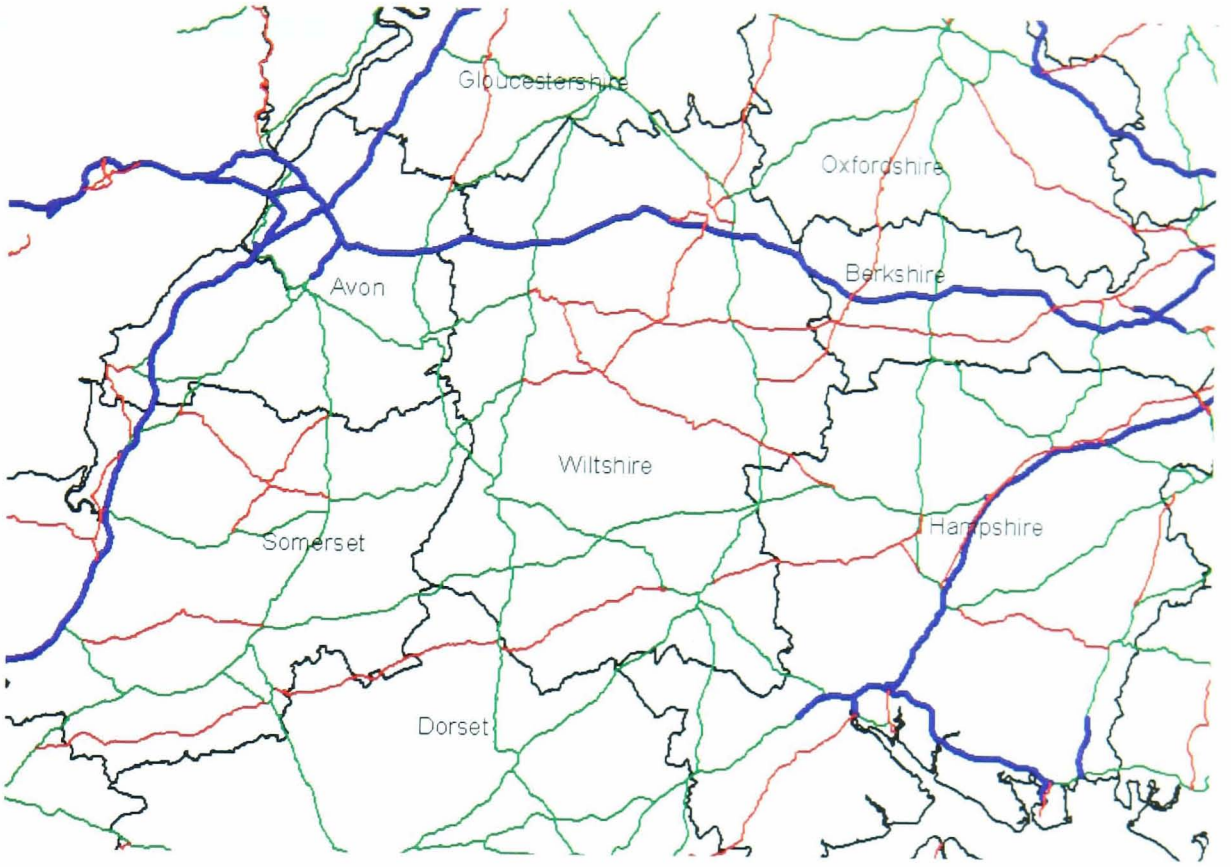

Figure 4.4: Map of part of Southern England in 1980's.

Green lines represent trunk roads. Red lines represent other A-roads other than trunk roads. Wiltshire lies in the middle of the map. Based on the common boundary rule, the first order neighours of Wiltshire are Oxfordshire, Berkshire, Hampshire, Dorset, Somerset, Avon and Gloucestershire. The spatial weights of these neighbours on Wiltshire can be determined by the road network. This can be done by observing the number of common roads passing between Wiltshire and its neighbours by different road types. For instance, considering the hierarchy, the function and traffic level of different road types, a motorway (dark blue) could be given the highest weight 2, a trunk road (green) could be weighted 1 and a A-road (red), not trunk road, could be weighted 0.5. Taking Berkshire and Oxfordshire as examples, there is one motorway (dark blue) crossing the boundary between Wiltshire and Berkshire, therefore, getting a weight 2. In addition, there are two A-roads (red), not trunk roads, crossing the two counties, weighting 1 (2×0.5). Therefore, the total weight of Berkshire on Wiltshire should be 3. There are only one trunk road (green) and one A-road (red) crossing the boundary between Wiltshire and Oxfordshire. The weight of Oxfordshire on Wiltshire should be $1 + 0.5 = 1.5$.

4.3 Univariate models with temporal effects

When longitudinal data are available, the response variable is denoted by y_{it}, which is the total number of accidents of a particular type at area or site i in time t. There are a number of ways to take account of temporal effects in the model. Based on model (4.2), temporal effects can be introduced by including a linear trend variable or by using different constant terms β_{0t} in different time periods.

Sometimes, temporal effects may have some specific structure such as temporal correlation. One way to model this correlation is to use an autoregressive model which introduces correlation between successive observations. Since one observation can only depend on previous observations but not future observations, a model that considers temporal correlation by using a first order autoregressive model can be formulated as:

$$\begin{aligned} y_{it} &\sim \text{Pois}(\lambda_{it}) \\ \log \lambda_{it} &= \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \\ \varepsilon_{it} &= \rho \varepsilon_{i(t-1)} + v_{it} \\ v_{it} &\sim \text{N}(0, \sigma_v^2), \end{aligned} \tag{4.4}$$

where ρ models the extent of correlation between successive observations.

There are some studies that model the temporal effects as a random walk (see Richardson et al., 2006). The random walk is formulated by a CAR prior. In order to formulate the CAR prior, the neighbours list needs to be identified. The first order neighbours for a period, excluding the first and the last period, are its previous period and the period next to it. The neighbour for the first period is the second period and the neighbour for the last period is its previous period.

4.4 Poisson model with spatio-temporal effects

A Poisson model with spatio-temporal effects can be developed by joining a model with temporal effects and a model with spatial effects. However, the formulation of the spatial

random effect using a CAR prior is more complicated because the data now cover several time periods. Suppose the neighbours list is fixed in time. For spatial random effects in different periods, the scale parameter will be $\tau_{\theta t}$. There are possibly two choices to decide it. One choice is to consider $\tau_{\theta 1} = \cdots = \tau_{\theta T} = \tau_\theta$, in other words, treat the scale parameters all the same in time. Under this circumstance, the spatial random effects can be specified using a CAR prior in two ways according to the following two assumptions. Firstly, the spatial random effect θ_{it} in area or at site i is constant over time. Secondly, the spatial random effect θ_{it} in area or at site i varies over time but the scale parameter for the CAR prior is constant. In order to predict accident count in area or site i in the future, the first assumption is required. Alternatively, different scale parameters for the CAR prior can be used for different years. This implies that the spatial random effects vary over time as well as the extent of spatial dependency.

When different scale parameters are used for the CAR prior in different periods, some temporal effects will be introduced in the model. However, these temporal effects do not describe the correlation structure in time. Therefore, temporal effects described in Section 4.3 can be included in model (4.3). For instance, if a linear trend variable t is included, spatial effects are assumed to be constant in time, and a first order autoregressive prior is chosen to model temporal random effects, then a model with both temporal effects and spatial effects can be written as

$$\begin{aligned} y_{it} &\sim \text{Pois}(\lambda_{it}), \quad \text{where} \\ \log \lambda_{it} &= \beta_0 + \beta_1 x_{it} + \delta t + \theta_i + \varepsilon_{it} \\ \theta_i | \theta_j \ [j \neq i] &\sim \text{N}\left(\sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}} \theta_j, \frac{\tau_\theta}{w_{i+}} \right) \\ \varepsilon_{it} &= \rho \varepsilon_{i(t-1)} + v_{it} \\ v_{it} &\sim \text{N}(0, \sigma_v^2). \end{aligned} \tag{4.5}$$

Both the temporal correlation and the spatial correlation in the residuals are expected to be removed after including both spatial effects and temporal effects in the model.

4.5 Multivariate models

All the models in the previous section are specified for only one type of road accident or for accidents of all types. If more than one type of accident are modelled jointly, say two, a Poisson model with log-normal random effects for the accident frequencies might be:

$$\begin{aligned} y_{1i} &\sim \text{Pois}(\lambda_{1i}) \\ y_{2i} &\sim \text{Pois}(\lambda_{2i}) \\ \log \lambda_{1i} &= \beta_{01} + \beta_{11} x_i + \varepsilon_{1i} \\ \log \lambda_{2i} &= \beta_{02} + \beta_{12} x_i + \varepsilon_{2i}, \end{aligned} \tag{4.6}$$

where ε_{1i} and ε_{2i} are assumed to be independent. This is a *multiple response model.* It can be noticed that the intercept term and the coefficient of the explanatory variable are different for different types of accidents. This model assumes that ε_{1i} and ε_{2i} are independent. However, in the same area i, one type of accident could be correlated with another type of accident. As described in Section 3.2, this type of correlation can be introduced by assigning ε_{1i} and ε_{2i} with a multivariate normal distribution $\text{N}_p(0, R)$. Here $p = 2$ because two types of accidents are considered here. The variance-covariance matrix R reflects the extent of correlation between different types of accident.

To include spatial structured random effects and unstructured random effects in a multiple response model, suppose there are only two types of road accidents and the Poisson means of them in each area i and in year t are λ_{1it} and λ_{2it}. The multivariate spatial model can be written as

$$\begin{aligned} \log \lambda_{1it} &= \mu_1 + \beta_1 x_{it} + \theta_{1i} + \varepsilon_{1it} \\ \log \lambda_{2it} &= \mu_2 + \beta_2 x_{it} + \theta_{2i} + \varepsilon_{2it}, \end{aligned} \tag{4.7}$$

where θ_{1it} and θ_{2it} are spatially structured random effects, and ε_{1it} and ε_{2it} are unstructured random effects.

One simple way to formulate these random effects is to treat θ_{1i} and θ_{2i} as independent and let them have different scale parameters τ_θ. The correlation between λ_{1it} and λ_{2it} can be introduced by the unstructured random effect ε_{1it} and ε_{2it}. This can be done by using a bivariate normal distribution, so that $\varepsilon_{1it}, \varepsilon_{2it} \sim \mathrm{N}_2(0, R)$, where R is the 2×2 variance-covariance matrix, which is $\begin{bmatrix} \tau_{\varepsilon 1} & \tau_{\varepsilon 12} \\ \tau_{\varepsilon 21} & \tau_{\varepsilon 2} \end{bmatrix}$ where $\tau_{\varepsilon 1}$ and $\tau_{\varepsilon 2}$ are the conditional variances of ε_1 and ε_2 respectively. The within-area correlation between ε_{1it} and ε_{2it}, is therefore $\tau_{\varepsilon 12}/\sqrt{\tau_{\varepsilon 1}\tau_{\varepsilon 2}}$. This is a measure of correlation in the two types of accidents.

The correlation between λ_{1it} and λ_{2it} can also be introduced in the spatial random effects θ_{1it} and θ_{2it}. A univariate CAR model can be extended to a multivariate CAR model in a number of ways (Jin et al., 2005), as introduced in Section 3.2. A straightforward approach is to use a multivariate normal distribution to handle the conditional distribution of the spatial random effects for different type of accidents. For model (4.7), suppose

$$\theta_{1it}, \theta_{2it} \sim \mathrm{N}_2\left(\left(\sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}}\theta_{1jt}, \sum_{j \in N[i]} \frac{w_{ij}}{w_{i+}}\theta_{2jt}\right), \frac{Q}{w_{i+}}\right)$$

where Q is the variance-covariance matrix for the spatial components.

The structure of Q can be expressed as $\begin{bmatrix} \tau_{\theta 1} & \tau_{\theta 12} \\ \tau_{\theta 21} & \tau_{\theta 2}. \end{bmatrix}$ Therefore, the within-area conditional correlation between the spatial components θ_{1it} and θ_{2it} is $\tau_{\theta 12}/\sqrt{\tau_{\theta 1}\tau_{\theta 2}}$ (Spiegelhalter et al., 2003). Combining both the multivariate spatial effects and the multivariate unstructured random effects, the conditional correlation between total random effects for the two types of accidents can be worked out. It is $(\tau_{\theta 12} + \tau_{\varepsilon 12})/(\sqrt{\tau_{\theta 1}} + \sqrt{\tau_{\varepsilon 1}})(\sqrt{\tau_{\theta 2}} + \sqrt{\tau_{\varepsilon 2}})$. Moreover, the shared-component models with a similar form of the model in 3.4 can be used to develop multivariate CAR models.

A multivariate model with spatio-temporal effects can be developed by adding a form of temporal effects, suggested in Section 4.3, in the model (4.7).

4.6 Model fitting and checking

All the models in this thesis were fitted in WinBUGS. Codes for selected models used in this research are included in Appendix F. More recently, WinBUGS can be called from R. This provides a more convenient way of model fitting and analysis of results. Other R packages used in this paper are 'spdep' (Bivand, 2004) and 'maptools' (Lewin-Koh and Bivand, 2004). In order to make the parameters reach convergence fast, all the explanatory variables after taking logarithms were standardized (by subtracting the mean and dividing by the standard deviation). For each model, two chains were simulated. For each chain, 10,000 to 20,000 iterations were generated. The last 2,000 iterations of each simulation chain for each parameter were kept for calculating the posterior mean. Under the simulation monitoring tool in WinBUGS, the convergence of the parameters can be visually examined. The $\hat{R}$ statistic introduced in Section 2.6.2 was used to confirm the status of convergence. $\hat{R}$ close to 1 indicates good convergence.

Methods introduced in Sections 2.6.3 and 3.3 can be used for model checking and comparisons. The Deviance Information Criterion (DIC) is usually used to compare the performance of models in different forms. A model with a lower DIC performs better. As a model becomes more complicated by taking account of spatial or temporal effects, the DIC is expected to decrease.

Moreover, examination of residuals helps to identify any problems with the model, for instance, the existence of temporal or spatial correlation. In the remaining part of the thesis, unless stated otherwise residuals used for model checking are *Pearson* residuals, defined as $(y_i - \hat{\lambda}_i)/\sqrt{\hat{\lambda}_i}$ (where $\hat{\lambda}_i$ is the posterior mean of λ_i). Moran's I statistic, which was introduced in the previous chapter, can be used to examine the spatial correlation in the residuals. A positive Moran's I indicates a positive correlation. The solution for this is to include appropriate spatial effects in the model, so borrowing information from the neighbouring areas. Positive and significant spatial correlation in residuals is likely to be obtained from models without any spatial effects. After accounting for the spatial dependency by including spatial effects, Moran's I in residuals is expected to drop and to become not significant. Upton and Fingleton (1985) state that the tests for residual

spatial autocorrelation are not valid when the model contains an autoregressive component. However, the models considered here contain an additional level of Poisson random variation. The residuals are defined as $(y_i - \lambda_i)/\sqrt{\lambda_i}$, where it is the model for the logarithms of the λ_i that contains the CAR component. In the absence of a more appropriate procedure, Moran's I is used here as an approximate indicator of the extent of spatial correlation in the residuals from the CAR models. The main problem of using Moran's I to examine spatial correlation in a Bayesian context is the replacement of λ_i by its posterior mean in the calculation of residuals. As mentioned in Section 2.6.3, in a Bayesian context, Bayesian residuals are more appropriate for model checking especially for examining spatial correlation in residuals. How Moran's I statistic should be obtained for residuals in a Bayesian context and be applied for model checking are introduced in the next section.

Moreover, residuals from the areal models can be plotted over the geographical map. Residual maps help to visualize concentrations of different ranges of residuals which thereby exhibit the influence of spatial correlation. These also show the progress in modelling the spatial correlation when spatial effects are included in the models. For instance, a map of residuals from a CAR model is expected to show fewer apparent clusters. In addition, for a CAR model, the posterior distribution of spatial random effects in an area or at a site can be obtained. An examination of such distribution will show whether an area or a site is associated with a positive (or negative) spatial effect. A map of spatial effects will help to identify areas with positive or negative spatial effects.

When models are fitted to longitudinal data, if no temporal effect is considered in the model, residuals from the model may display temporal autocorrelation. In order to examine the temporal correlation in the residuals, the residuals need to be separated into T groups, where T is the total number of years. Then the Pearson correlation coefficient can be worked out to check the extent of temporal correlation in the residuals. If the correlation is high, the indication could be there is temporal correlation in the data and this should be considered in the model. In this context, spatial correlation needs to be examined in the residuals for different time periods. Models with appropriate forms of

temporal effects are expected to reduce the temporal correlation in the residuals compared with models without including any temporal effect.

4.7 Posterior distribution of Moran's I

As explained earlier, if there are N observations used to fit a Poisson regression model with means λs and M simulations are saved for model estimation, then we will have a $N \times M$ matrix of simulated values for λs, with element $\lambda_i^{(j)}$ representing the estimate of Poisson mean for area or site i in the jth simulation. Therefore the Bayesian residuals also form a $N \times M$ matrix, with element $\frac{y_i - \lambda_i^{(j)}}{\sqrt{\lambda_i^{(j)}}}$. For residuals from each simulation, the value of Moran's I can be calculated. Therefore, we will have M values of Moran's I that compose a posterior distribution of I. If there is a significant level of positive spatial correlation in Bayesian residuals, most values of I should be positive and large enough to be statistically significant. This is a better approach to examining the spatial correlation in residuals than using residuals calculated from the posterior mean of λ_i, because only residuals obtained from λ_i in the same simulation will completely reflect the real spatial pattern in residuals.

Gelman et al. (2000) proposed a general approach for model checking using posterior predictive simulations. Let y denote the observed data and $y^{(j)}$ denote predicted values from a model based on a vector of parameters β in the jth simulation. A statistic $T(y, \beta)$ can be chosen to compare the difference between realized and predicted values with the statistical significance of the test summarized by a p-value, $p = \Pr(T(y^{(j)}, \beta) > T(y, \beta)|y)$. This is a one-sided p-value. A two-sided p-value will be $2\min(p, 1-p)$. If M simulations are used, p-value can be summarized by $\sum_{j=1}^{M} = H_{j|T(y^{(j)},\beta)>T(y,\beta)}/M$, where $H_j = 1$ if the condition is true and 0 otherwise. For a two-sided test, a very small p-value, say smaller than 0.05, will indicate large and systematic differences between realized and predicted values and therefore a misfit of the model to the data.

Gelman's approach explained above can be used for model checking by using Moran's I statistic as the $T(y, \beta)$ statistic. We will calculate a p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$, where $\lambda^{(j)}$ are fitted Poisson means in the jth simulation and $I()$ calculates the value of

Moran's I in the residuals. If there is a significant level of positive spatial correlation in Bayesian residuals, the posterior distribution of Moran's I should contain enough number of significant values of I. However, the I statistic for residuals $\frac{y^{(j)}-\lambda^{(j)}}{\sqrt{\lambda^{(j)}}}$, where $y^{(j)}$ are simulated as $y^{(j)} \sim \text{Pois}(\lambda^{(j)})$, should output small values of Moran's I that are expected to be nonsignificant. In such a case, a very small p-value will be obtained and will suggest that the model is not fitted well and spatial correlation needs to be taken account of in the model by an appropriate approach. On the contrary, if the spatial correlation has already been considered in a model, the posterior distribution for values of Moran's I in residuals based on true values y should have most values to be small and nonsignificant. So as for the $I()$ statistic for the predicted values $y^{(j)}$. Therefore, by comparing two posterior distribution of I, a p-value around 0.5 would suggest a proper fit of the model.

The above method is proposed based on Gelman's approach for diagnostic check using posterior predictive simulations. How well this method works for model checking using Moran's I statistic in this research is examined in the latter chapters. However, for simplicity, examining the value of Moran's I in Pearson residuals based on the posterior mean of λ_i is still used as a main approach for investigating the existence of spatial correlation in the residuals.

Chapter 5

Methodology: Variables and data

Statistical models for road accidents can be used to explain the variation in accident data, study the relationship between the number of road accidents and variables that describe traffic levels, road characteristics and other relevant factors and predict the accident frequency in the future. Therefore, it is important to decide how to choose the explanatory variables that should be included in a model. The choice of explanatory variables depends on both the understanding of the factors that contribute to road accidents and the availability of the data. In practice, data are not available for all the desired variables, and even if they are available, their quality are not always good.

In this chapter, some existing problems with accident data and the choice of the response variables are discussed first. Then, what types of variables need to be included in accident models and how they can be measured are explained. Later, how explanatory variables were chosen and how the data were collected in this research are introduced.

5.1 Accident data and response variables

The main source of accident data in the UK is STATS19. Section 2.2 has given a general introduction to the STATS19 data. As shown earlier, using statistical models to analyse accidents, accident data need to be aggregated over space. To be more specific, for areal models, they need to be aggregated at the area level like local authority; for network models, accidents at junctions or on links need to be aggregated. Whether the true number

of total accidents in an area or at a site can be obtained from the STATS19 data partially depends on the accuracy of the location information in STATS19. However, location information like the grid references was sometimes wrongly recorded. This may cause a problem for spatial aggregation of accidents, especially when the total number of accidents at a junction or on a link needs to be obtained. Moreover, accidents involving no personal injury are not recorded in STATS19. Therefore such accidents cannot be taken account of in the analysis.

Most accident models in the literature use the accident count in an area or at a site as the response variable. There are some studies that modelled casualties at the area level rather than the accident counts (see, for instance, Noland and Quddus, 2004). In disease mapping, the response variable is usually the number of persons that have a certain kind of disease. Each incidence of the occurrence of the disease is associated with only one individual. In the context of road accidents, each accident may involve a number of injured individuals. This creates an internal link among the casualties from the same accident. These casualties are not independent. Therefore, areal models for road casualties are more complicated. It is more straightforward to model accident counts and to identify factors contributing to high accident frequencies. In this research, the response variable is the number of a particular type of accident at an area, such as a local authority or a ward, or the number of accidents at a site, such as a junction or a link, during a period of time.

5.2 Choice and measurement of explanatory variables

As explained in previous chapters, accident data, such as STATS19, need first to be aggregated by location and time in order to fit statistical models. For areal models, such as models for local authorities, the response variable is the total number of road accidents in an area during a fixed period. Therefore all the explanatory variables should be available at the same geographical level. Similarly, for accidents on links or junctions, the explanatory variables should be available for the corresponding spatial unit. According to the different type of factors that contribute to accidents, the explanatory variables can be grouped into several categories. They are now introduced in turn.

5.2.1 Traffic and road characteristics

Road accidents occur because there is traffic on the roads. Therefore, two most obvious factors that contribute to road accidents are exposure to traffic and road characteristics. Exposure to traffic depends on traffic levels that are often measured by annual average daily flows (AADF), which is used in models for accidents on links or junctions, or annual traffic volume (vehicle-km), which is often used in areal models. The Department for Transport website explains how national traffic estimates are made (see Department for Transport, 2004). The calculation of both AADF and traffic volume uses information from manual and automatic traffic counts (Department for Transport, 2004). For major roads (motorways and A-roads) in England, the traffic on most links is manually counted at a statistically random point at regular intervals. This is done mostly in so-called 'neutral' weeks, namely most weeks in March, April, May, June, September and October, avoiding main holiday periods. For minor roads, complete coverage of the road network is not practical. Minor roads can be grouped into three road classes, namely class B, class C and unclassified roads. Roads can also be categorised as urban and rural roads. Urban roads are defined as those within the boundaries of the Urban Area polygons for settlements of 10,000 population or more, based on the 2001 Population Census (see Department for Transport, 2004). On the outskirts of urban areas, bypasses are normally treated as rural even if part of the road may lie within the urban area polygon. Conversely, roads between urban areas with short lengths outside the polygons are normally treated as urban. However, before 8 May 2003, roads were instead classified as built-up and non built-up. Built-up roads were those with a speed limit of 40mph or less. Non built-up roads were those with a speed limit higher than 40mph.

For each of these road types, the average flow is measured by carrying out a number of counts along them. One limitation of manual counts is that the traffic is counted for only 12 hours on each visit. Thus the counts give no information about traffic at night, at weekends, over public holiday periods, and little information about the non-neutral months. Therefore, to get reliable estimates of traffic flow, data obtained from automatic counters are used. For both the manual count and the automatic count, the traffic is

counted by different vehicle types, such as car or bus.

An estimate of the AADF at a site is then calculated by multiplying the manual count data at the site by factors derived from the automatic counts on similar roads in the same year. The annual traffic estimates for a major road or for roads of a particular category in each local authority can be obtained after taking account of the road length for the relevant road category. For instance, a major road link of length 2km with an AADF of 50,000 has a traffic volume of 100,000 vehicle-kilometres ($2 \times 50,000$ per day). This equates to 36.5 million vehicle-kilometres a year.

Road characteristics are important factors related to accident frequencies. As explained above, in order to obtain the annual traffic estimates, road lengths of different types of road are needed. Moreover, the road length itself is often included in a model as an explanatory variable. Traffic levels and the density of the road network can be very different for different road types. Therefore, lengths of roads are often disaggregated by factors like road class. Other road characteristics include the curvature and the gradient of the road, number of carriageways, number of lanes in the road, road width, etc.

The AADF can be used directly in models for junctions and links on major roads. But for areal models, the annual traffic estimates are often used. The choice of the level at which the road lengths and the annual traffic estimates should be disaggregated depends on the attributes of the response variable. In other words, it depends on the level at which the accident data are disaggregated. For instance, if the response variable is the number of the accidents on urban A-roads in a local authority, ideally, the traffic variables need to be traffic estimates for urban A-roads, possibly together with traffic estimates of other types of roads, and the road length variables need to be lengths of urban A-roads, perhaps together with lengths of other types of roads.

The complexity of a road network can be measured by the number of junctions per unit road length. A junction can be a roundabout, a T-junction, a crossing or other types. Theoretically, more junctions lead to more traffic conflicts, in turn leading to greater risk of road accidents. Therefore the junction density in each local authority can be used as an explanatory variable in the areal model.

5.2.2 Proxy variables for traffic

Traffic estimates such as AADF or annual traffic estimates cannot measure traffic perfectly. Especially for small areas, such as wards, traffic levels are difficult to obtain. Using proxy variables for traffic, which are available at the desired geographical level, can be a solution. Resident population and employment are two general forms of proxy variables for traffic. The more population and employment, the more activities and trips are generated on roads. For instance, Bailey and Hewson (2004), and Noland and Quddus (2004) use employment and resident population as proxies for traffic levels. Moreover, the population by mode of travel to work could be even better proxies for the traffic since they include more information about the composition of the local traffic.

5.2.3 Characteristics of the geographical area

For areal models, the characteristics of the area are often considered. The economic conditions and extent of urban development can be different for metropolitan districts and more rural areas. Both of these variables, such as unemployment rate and housing density, can be indicators of road conditions and traffic characteristics. Other data relevant to the characteristics of the geographical area includes how large it is.

Some studies (for instance, Noland and Quddus (2004)) use variables like population density and road density rather than total population and road length. When the Poisson log-linear model as described in model 2.2 is used, including the two density variables is a special case of including three variables—area, population and road length expressed in logarithmic form.

5.3 Data collection and preparation

Variables that are often included in the areal model have been discussed in the previous section. Three areal datasets have been used to fit the spatial models proposed in the last chapter. The first dataset is from a previous research project (see Jarrett et al., 1989). These data were used to fit models with spatial random effects at the local authority level.

The second dataset includes accident data for a five year period and other variables most of which cover the same period. These data were used to fit the multivariate spatio-temporal models at the local authority level. The third dataset consists of accident data and other data at the ward level in the West Midlands. It was used to fit the multivariate CAR models. Two sets of data were used to fit models for accidents on a road network. Some of the geographical data, for instance the boundary map and the road characteristics, were restructured and prepared in ArcView 3. Other data were structured and transformed in R, SPSS and Excel. Details about how these data were obtained and prepared are explained now.

5.3.1 Data for local authorities in England from 1983 to 1986

In order to develop spatial models for road accidents at the local authority level and to investigate how models with spatial effects can improve conventional models that do not take account of spatial effects, some data used in a previous research for TRL by Jarrett et al. (1989) were used. These data include accident data in England from 1983 to 1986 at the local authority level. There were 108 geographical units during that period, which were 39shire counties, 36 metropolitan districts in 6 former metropolitan counties, and 33 London boroughs. Table A.1 in Appendix A gives a full list of names of the local authorities. The accidents were disaggregated according to severity, road class and speed limit. Only accidents on built-up A-roads were used to fit the models developed in this thesis. Here, built-up roads are defined as roads with speed limit 40 mph or less. The original data include other variables, such as traffic, road length and population. In this study, for models of accidents on built-up A-roads, five variables were chosen to be included in the models. Since the response variable is the number of accidents of different severity on built-up A-roads, explanatory variables regarding to road length and traffic volume are for built-up A-roads only. Other explanatory variables include population, geographical area and number of licensed vehicles.

In order to construct the neighbours list for the CAR models and explore the spatial distribution of residuals, a geographical boundary map of England was created in Ar-

cView. Boundary maps of England by different administrative division during different periods can be downloaded in an ArcView-readable form from UKBORDERS (EDINA, 2007). The boundaries of metropolitan districts and London boroughs were obtained based on the district map in 1991. These were combined with the boundaries of other counties that were extracted from the county map in 1991.

To take account of the fixed spatial effects in the models (see Section 4.2.1), two types of factors were used. The first type of factor was represented by two dummy variables that describe whether a local authority is a shire county, a metropolitan district or a London borough. Another type of factor was represented by 7 dummy variables that describe which metropolitan county (including London) a borough or a metropolitan district belongs to, namely London, Great Manchester, Tyne and Wear, West Yorkshire, South Yorkshire, Merseyside, and West Midlands.

In order to model structured spatial effects by a CAR prior, the spatial neighbours list and the spatial weights are needed. Based on the discussion of constructing a neighbours list and a weighting choice in Section 4.2.3, three different spatial neighbours lists and three different weighting schemes were used for this dataset. For the neighbours list, the first choice is based on the definition of neighbours as local authorities that share at least one common boundary. For the second choice, local authority j is defined as a neighbour of local authority i if there is at least one common boundary between them, and at the same time at least one of the following conditions is satisfied: (a) there is at least one common motorway going through i and j; (b) i and j are in the same metropolitan county (including London) and there is at least one common trunk road through them. The last choice is slightly different from the second one. In this, local authority j is defined as a neighbour of local authority i if there is at least one common boundary between them, and at the same time there is at least one common motorway or trunk road through local authority i and j.

Three different weighting schemes were used. For the first type of neighbours list,a 1–0 scheme and weights defined by Euclidean distance are both used. For the second type of list, only a 1–0 scheme is used. For the third list, the weights are determined by the

number of common roads that crossed the neighbouring local authorities as explained in Section 4.2.3.

5.3.2 Data for local authorities in England from 2001 to 2005

This dataset consists of data covering a five year period. During the 1990s in England, some cities, large towns and groups of neighbouring towns became unitary authorities. Unitary authorities in England are typically defined as any authority which is the sole principal council for its local government area (see Secretary of State for the Environment, 1994). Therefore, a boundary map that is different from the one used for the first dataset needs to be created when modelling the number of road accidents at the local authority level from 2001 to 2005. Based on boundary maps of counties and districts in 2001 obtained from UKBORDERS (EDINA, 2007), a boundary map of local authorities was created in ArcView. 149 geographical units were identified, including shire counties, metropolitan districts, unitary authorities and London boroughs. Table A.2 in Appendix A gives a full list of names of the local authorities. One problem with the district boundary map in 2001 is that the map does not show the rivers that are available in the district map in 1991. This can influence the result of the neighbours list. If a river separates two local authorities and there are no roads crossing the river between the two local authorities, they are not considered as neighbours. However, using a boundary map without river lines, these two local authorities appear to have a common boundary and be neighbours. Therefore, the boundary map of local authorities in 2001 was modified according to the administrative area boundaries and river lines obtained from Meridian 2 which is a type of Ordnance Survey datasets (see Meridian^{TM}, 2007).

Five years' accident data were obtained from STATS19 (Department for Transport, 2007*b*). Based on the 149 local authorities identified in the boundary map, the local authority code variable in STATS19 was recoded in order to aggregate accidents at the right geographical level. The total accidents in each local authority were disaggregated by accident severity. They were not disaggregated by road class because, for the traffic variable, only total traffic estimates are available. Both the traffic data and the road length

data were obtained from the Department for Transport. Traffic data measured by vehicle-kilometres are available for both the car traffic and all vehicle traffic from 2001 to 2005 (Department for Transport, 2007*d*). Therefore, two variables were chosen to describe traffic, namely the traffic for cars and the traffic for other vehicles. The second one was obtained by subtracting the traffic for cars from the traffic for all vehicles.

The Department for Transport published the data of road length at the local authority level for different road categories. However, the data are available only for the year 2004. Therefore they were used for all the 5 years, which is a limitation of the data. According to how the road length is disaggregated by road categories in the original dataset, three variables were constructed. The length of A-roads is the sum of the lengths for the rural trunk roads, urban trunk roads, principal urban roads and principal rural roads. The length of B-roads is the sum of the lengths for the rural B-roads and urban B-roads. The length of other minor roads is the sum of the lengths for the rural and urban C-roads and rural and urban unclassified roads.

As explained earlier, the number of junctions in an area can be an indicator of the complexity of the road network and a measurement of the number of potential traffic conflicts. The total number of all types of junctions in each local authority was obtained by counting the number of nodes, corresponding to junctions, obtained from Meridian 2 (Meridian^{TM}, 2007).

Other variables are geographical area (in km^2) and the population (in thousands). Both of them were obtained from the National Statistics website (see Office for National Statistics, 2001). Population estimates are available for the whole study period.

Two types of factors were used to take account of the fixed regional effects. Definitions of these factors are the same as those used for data in the 1980s introduced in the previous subsection. For the CAR model, only one type of neighbours list was used, which defines neighbours based on the condition of sharing at least one common boundary.

5.3.3 Data for wards in the West Midlands in 2001

This set of data includes accident data and other relevant variables in the West Midlands. The metropolitan areas of the West Midlands consist of 7 districts, namely Birmingham, Coventry, Dudley, Sandwell, Solihull, Walsall and Wolverhampton. Due to the limitation of the availability of data for other variables, only accidents in 2001 were considered. These data were used to develop multivariate CAR models.

Most of these data were obtained from the SPECTRUM database developed by Mott MacDonald (SPECTRUM, 2007). SPECTRUM is a web-based geographic information system (GIS) supplying road, traffic, accident and census data of the 7 Metropolitan Borough Councils in the West Midlands. The geographic and road data in SPECTRUM come from products of the Ordnance Survey (Ordnance Survey, 2007), including Boundaryline, MasterMap, and Ordnance Survey Centre Alignment of Roads (OSCAR) Asset Manager. Based on boundary lines of census wards in 2001, there are in total 162 wards in the West Midlands. Table A.3 in Appendix A gives a full list of names of the wards.

SPECTRUM Accidents is a module within the SPECTRUM core system. Accident data in SPECTRUM are based on the 'Collision Report' forms collected by the West Midlands police. They include all the information available in the STATS19 data. One difference between the two collecting systems is the precision of the accident location. In the STATS19 form, the location of an accident is recorded by a 10-digit grid reference with 5 digits for the easting reference and 5 digits for the northing reference. The fifth digit of northing and easting defines a 10-metre unit. The West Midlands 'Collision Report' uses a 12-digit grid reference which locates accidents at the 1-metre level. This may locate road accidents on a road network more precisely. However, when an accident occurs, it is difficult to decide the location of it as precisely as 1 metre because even one vehicle is more than 1 metre long. The number of road accidents with a particular type in each ward can be obtained by using an aggregating function in SPECTRUM. They can then be disaggregated within each ward by severity.

SPECTRUM contains data for major roads in the West Midlands, including motorways, A- and B- roads based on the OSCAR data, a product of the Ordnance Survey. The

OSCAR data are based on vector data in a link-node structure. For OSCAR data, a road is made up from a series of links and a link can be a section of road between two adjacent junctions or changes in carriageway type. Links are defined by road centrelines with some additional attributes attached. The attributes include road name, road number, length of the link, and road type, for instance, dual carriageway, single carriageway or roundabout. In this study, OSCAR data for different classes of roads were exported from SPECTRUM and imported into ArcView. For minor roads, including C-roads and unclassified roads, the data were obtained from other layers of roads in SPECTRUM. The total length of the roads in a ward was calculated by adding up lengths of all the links within the ward by road class. There are some links that cross the boundaries or neighbouring wards. They were split at the intersection between the link and the boundary line in Arcview. Three explanatory variables were constructed for the length of roads. They are length of A-roads, B-roads and minor roads. A junction is defined by a node in OSCAR. Unfortunately, node data are not available in SPECTRUM. Therefore, node data for junctions in the West Midlands were extracted from Meridian 2. The number of junctions in each ward was calculated by counting the total number of nodes within a ward polygon.

Other explanatory variables were extracted from the 2001 Census at the ward level. The area of a ward is measured in hectares. The average size of the wards in the West Midlands is about 555 hectares, with the smallest being 167.7 hectares and the largest being about 5836.6 hectares.

One limitation of the West Midlands data is that the traffic volume data are not available at the ward level. Therefore populations travelling to work by different transport means were used as proxies for traffic. Four variables were chosen from the Census data. They are population travelling to work by car as driver, population travelling to work by car as passenger, population travelling to work by bus, population travelling to work on foot. These variables were selected because all these travel means will generate traffic on roads.

Moreover, the West Midlands consists of 7 metropolitan districts. In order to examine the effects of the districts, a factor represented by 6 dummy variables was constructed to

be included in the models with fixed regional effects. For the CAR model, the neighbours list is still determined by the condition of boundary sharing. However, both first order neighbours and higher order neighbours are considered for the ward data because a ward is relatively small. When higher order neighbours are considered, the weights are chosen to be inversely proportional to the Euclidean distance between the centroids of the neighbouring wards.

5.3.4 Data for the M1

The M1 extends approximately in a north-south direction. It passes through the outer London, East of England, East Midlands and Yorkshire & the Humber traffic regions. Figure 5.1 plots the layout of the motorways in England with the M1 plotted in black.

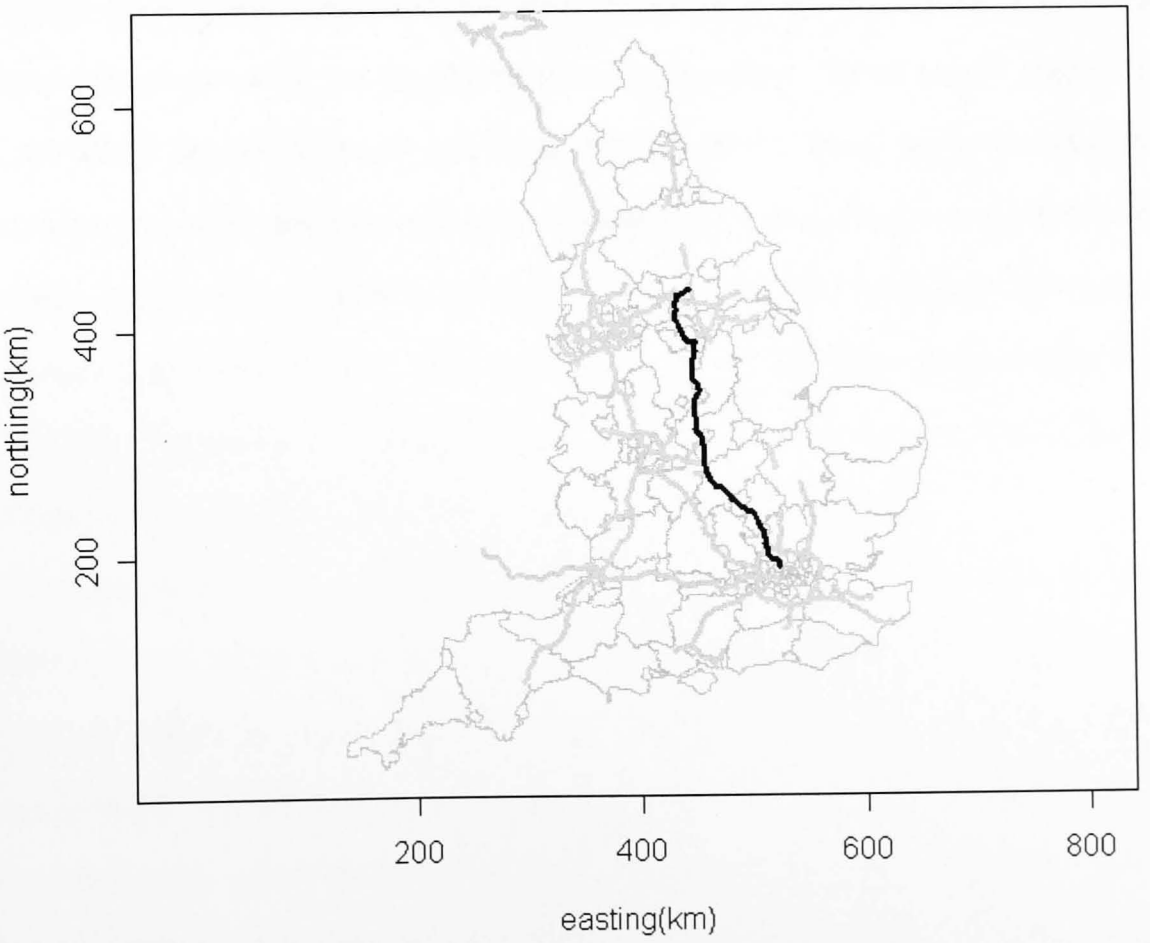

Figure 5.1: Layout of the motorways in England.

Accident data were obtained from STATS19. Accidents on the M1 from 1999 to 2005 were selected in SPSS by using variables of 1st road class and 1st road number. Accidents that are coded as junction accidents in STATS19 were excluded from the analysis. The accident data were imported into the geographical information system (GIS) software ArcView in a separate layer in addition to the other layers including the local authority boundary map, the M1 road, other motorways (obtained from Meridian 2).

The traffic flow data were obtained from an online traffic database (Department for Transport, 2006*a*). This traffic database provides annual average daily flow (AADF) of traffic for all major roads in UK. In this the roads are broken up into a series of links. Each link comprises a stretch of major road between two consecutive junctions with other major roads. A link may also start or end at a local authority boundary or an urban/rural area boundary. A traffic count takes place on each link of the major road network. The variables in the data include a unique reference for the road link and the grid reference for the traffic count point. Therefore, the data can be imported to ArcView and the location of the traffic count point can be plotted in a separate layer. There are 77 observations of the AADF for the M1 in every year from 1999 to 2005. There are some adjacent links that are separated by the local authority boundary. However, there is no junction between the links, therefore the AADF is the same for these links. For such links, they are treated as a single link in the analysis. After combining this type of link, 59 links were identified on the M1. Therefore, the 'shapefile' of the M1 was edited in ArcView to produce 59 equivalent links, each of which was assigned with a unique link ID.

Accident data were aggregated for each link by using the 'Spatial Join' option of the 'GeoProcessing' extension in ArcView. For each point data, this option assigns the ID of the nearest link and the nearest distance to it. Therefore, the point data can be aggregated by using the ID variable. However, for some accidents, their distance to their nearest link is extremely high. After checking these accidents, it was found that there are errors in the location variables, measured by grid reference, for them. For some of them, the location variables have missing values. Some of them have zero for the locations while some of them have only 4 digits that are clearly wrong. For these accidents, the following criteria

were set to assign them to a correct link ID or to exclude them from the data. There is a variable in the accident data for the local authority code. This variable was used to check the correct local authority that an accident occurred. If there is only one link of the M1 in this local authority, the ID of this link was assigned to this accident. If there are two or more links of the M1 in this local authority, the accident was excluded from the data. If there is no link of the M1 in the local authority, the accident was excluded from the data. Many of these accidents with incorrect location records are found to be in the South Yorkshire according to the local authority code.

It is relatively easy to identify the structure of neighbours for a single long road like the M1. As described in Chapter 4, a road network can be represented by a node-link graph. Figures 5.2 presents the M1 in a node-link graph. Each node in the graph corresponds to a motorway junction. Each link between two nodes represents a road link. 59 links have been identified. Each link is numbered from the left to the right by a link ID using numbers 1 to 59. There are 4 spurs from the main road each represented by a vertical or a sloping link in the plot. The length of each horizontal link is scaled based on the true length of the road link, but not for the spurs because they are short and cannot be visually shown if the same scale ratio is applied. The axis at the bottom measures the length of the M1. The total length of the M1, excluding the four spurs is about 307 km.

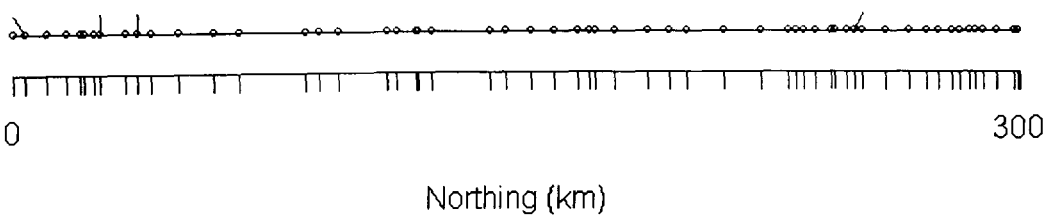

Figure 5.2: Node-link graph for the M1.

Figures 5.2 shows that most links have two neighbours and the link on the far right has only one neighbour. The 'sloping' links join the main motorway at restricted junctions. Therefore, the first spur is a neighbour only of the link on its right and the fourth spur is

a neighbour only for the link on its left. The two 'vertical' links are neighbours for both the links on their left and the right because their traffic can join or leave from both sides. Therefore, the links that are next to a vertical link have three neighbours and each vertical link has two neighbours.

5.3.5 Data for junctions in Coventry

Data used to analyse accidents at junctions were obtained from the SPECTRUM database which was introduced in Section 5.3.3. Coventry was chosen to be the study region because its road network is not as complicated as that of Birmingham but still shows enough level of complexity. The study network consists of A- and B-roads, with junctions the intersections of these roads. Data about roads and accidents at major junctions from 2002 to 2006 were downloaded from SPECTRUM and imported to ArcView in a number of layers. There are 30 roundabouts and 25 junctions of other types. Positions and shapes of roundabouts in Coventry are available in the road data. Therefore, roundabouts were identified first. For each roundabout, it has a unique ID which is consistent with the junction ID variable in the accident data. Therefore, the total number of accidents at a roundabout is easy to obtain by matching the ID variables in the data of roundabouts and accidents. However, for other types of junctions, there is no such ID variable. Since a junction accident is defined as an accident occurring within 20 metres from a junction, for junctions of other types, the total number of accidents at each junction was obtained by searching all the point events that represents locations of accidents within 20 metres from the junction and getting the total count.

In order to study the spatial correlation in the junction accidents, neighbours of a junction needs to be defined. Figure 5.3 is a node-link graph displaying the connectivity between junctions. It shows an approximate illustration about how major junctions spread out over the road network in Coventry. Each node represents a junction. A link between two nodes means there is a road joining the nodes. Nodes at each end of a link are regarded as neighbours. The spatial weights that reflect the extent of spatial dependence were chosen to be inversely proportional to the length of the road sections between the

neighbouring nodes.

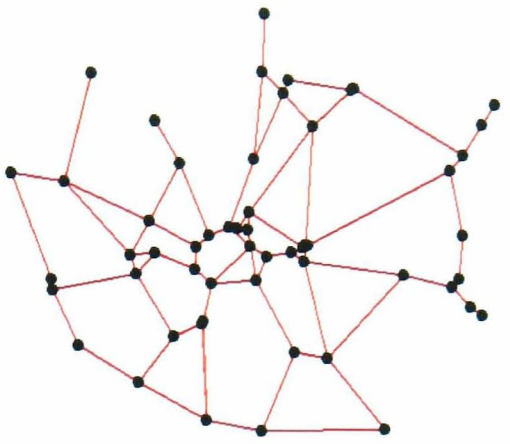

Figure 5.3: Neighbours structure of major junctions in Coventry.

Although the SPECTRUM database provides traffic count data for some locations in Coventry, the counts were made at different times in different locations. Traffic was counted before 2000 in some of the locations while in other places it was counted after 2000. Therefore it is difficult to obtain the traffic levels that can be used as an explanatory variable in the junction models. Moreover, traffic levels at a junction depend on the traffic movement on all the arms of the junction. However, locations of the traffic counts are relatively sparse in Coventry. Therefore only one explanatory variable, namely the junction type, was used. A junction can be a roundabout, a crossing or a T junction. The road data from the SPECTRUM provides information for the roundabouts. For other junctions, the types of them were identified by observing the road layouts in ArcView. A categorical variable represented by two dummy variables was used to describe the type of a junction.

Chapter 6

Areal models for accident frequencies

Reasons for the need to include spatial random effects in the conventional accident models have been discussed in Chapter 2. Based on the examination and the comparison of several possible approaches to taking account of such spatial effects, the method to develop spatial accident models has been proposed in Chapter 4. Such models also consider temporal effects and the correlation between accidents of different severity. This chapter aims to explain how these models are fitted using some real datasets. Details of the datasets were described in Chapter 5. The first two datasets contain data at the local authority level and cover several years. For these data, models with both spatial effects and temporal effects are studied. Moreover, multivariate models that jointly model accidents of different severity are developed for the second dataset. The last dataset is used to fit multivariate CAR models at the ward level. Results of models using different datasets will be shown in turn, including the comparisons of different forms of models and the interpretation of the results.

6.1 General description

The approach adopted to develop and extend models is to start from a simple model and then make it more complicated by including extra terms in the models. Details of different types of models used in this research have been introduced in Chapter 4. Recall the second

line in the univariate model (4.5) in Section 4.4.

$$\log \lambda_{it} = \beta_0 + \beta_1 x_{it} + \delta t + \theta_i + \varepsilon_{it} \tag{6.1}$$

This model is extended from a Poisson log-linear model that includes only the explanatory variable. In addition to the explanatory variable x_{it}, it includes a linear time trend t, a spatially correlated random effect θ_i and a random effect that is spatially independent but temporally correlated. When fixed spatial effects instead of random spatial effects are assumed, θ_i will be replaced by $\sum_{L=1}^{M-1} \alpha_L D_{iL}$ (see Section 4.2.1). Moreover, based on the univariate model, multivariate models can be built-up. Two forms of multivariate models are fitted in this chapter. They are multivariate CAR models and shared component CAR models. Details of these models have been explained in Sections 4.5 and 3.2.

By gradually introducing spatial effects and temporal effects in the Poisson log-linear model, the influence and the importance of including such effects can be examined. This can be achieved by a number of ways. For instance, comparing the DICs; checking the existence of spatial or temporal autocorrelation in Pearson residuals. Pearson residual is defined in this thesis as $(y_{it} - \hat{\lambda}_{it})/\sqrt{\hat{\lambda}_{it}}$, where $\hat{\lambda}_i$ is the posterior mean of λ_i (see Section 3.3). Moreover, the estimates of the coefficients in equation (6.1) can be studied. For instance, a positive median or mean of β_1 indicates a positive effect of the explanatory variable x_{it} on the expected number of accidents λ_{it}. In other words, a local authority that has a larger x_{it} will have a higher expected number of accidents compared with other local authorities that have similar values for other explanatory variables. In addition, in a Bayesian context, a p-value is seldom used to make a conclusion of whether an explanatory variable is significant or not. A credible interval, normally 95%, is usually preferred to examine the contribution of an explanatory variable in a model. A general rule is to check whether the credible interval includes zero. If it does, especially when the median is close to zero, this indicates that the explanatory variable does little to explain the variation in the response variable.

In order to make comparisons of different models more convenient, a unique name will be given to a model with a specific structure that describes what terms are included at

the right hand side of equation (6.1). How a model is named, depending on its structure, is explained below. Based on the simplest form of a Poisson log-linear model denoted by PL, if appearing in a model name, 'fe' means the inclusion of the fixed effects that describe the type of a local authority (shire county, metropolitan county or London borough); 're' corresponds to the inclusion of the metropolitan county effects; 'N' means log-normal random effects are included; 'tr' means a linear time trend variable is included; 'temp' indicates the inclusion of random temporal effects. For instance, model PLNre will correspond to a Poisson log-linear model with log-normal random effects and metropolitan county effects. When a CAR (conditional autoregressive) model is used, the name of the model will start from either ICAR (the intrinsic CAR which includes only spatially structured random effects) or CCAR (the convolution CAR which includes both spatially structured and unstructured random effects) and may include a footnote that describes how the neighbours list and the spatial weights are defined.

6.2 Models for accidents at the local authority level in England from 1983 to 1986

The dataset used to fit the models here was obtained from previous research. Details of the data have been introduced in Section 5.3.1. Five explanatory variables are included in the models. They are area, population, length of built-up A-roads, traffic volume on built-up A-roads and number of licensed vehicles. In a Poisson log-linear model, the logarithm of λ_i, the expected number of accidents in a local authority, is linked to a linear combination of the explanatory variables in logarithmic forms. A scatter plot matrix for all the variables in logarithmic forms will help to examine the relationships between these variables. Figure 6.1 illustrates the relationship between the response variables (for accidents of different severity) and selected explanatory variables in 1986. Such plots for the other years show similar results therefore are not included here. Figure 6.1 suggests three implications. Firstly, the explanatory variables are positively correlated with the response variables—counts of a particular type of accident. Secondly, the explanatory variables are

positively correlated. Finally, the response variables are also positively correlated. Moreover, an outlier is observed at the left bottom corner of most scatter plots. It corresponds to the city of London, where values of the explanatory variables were the smallest in the 108 local authorities.

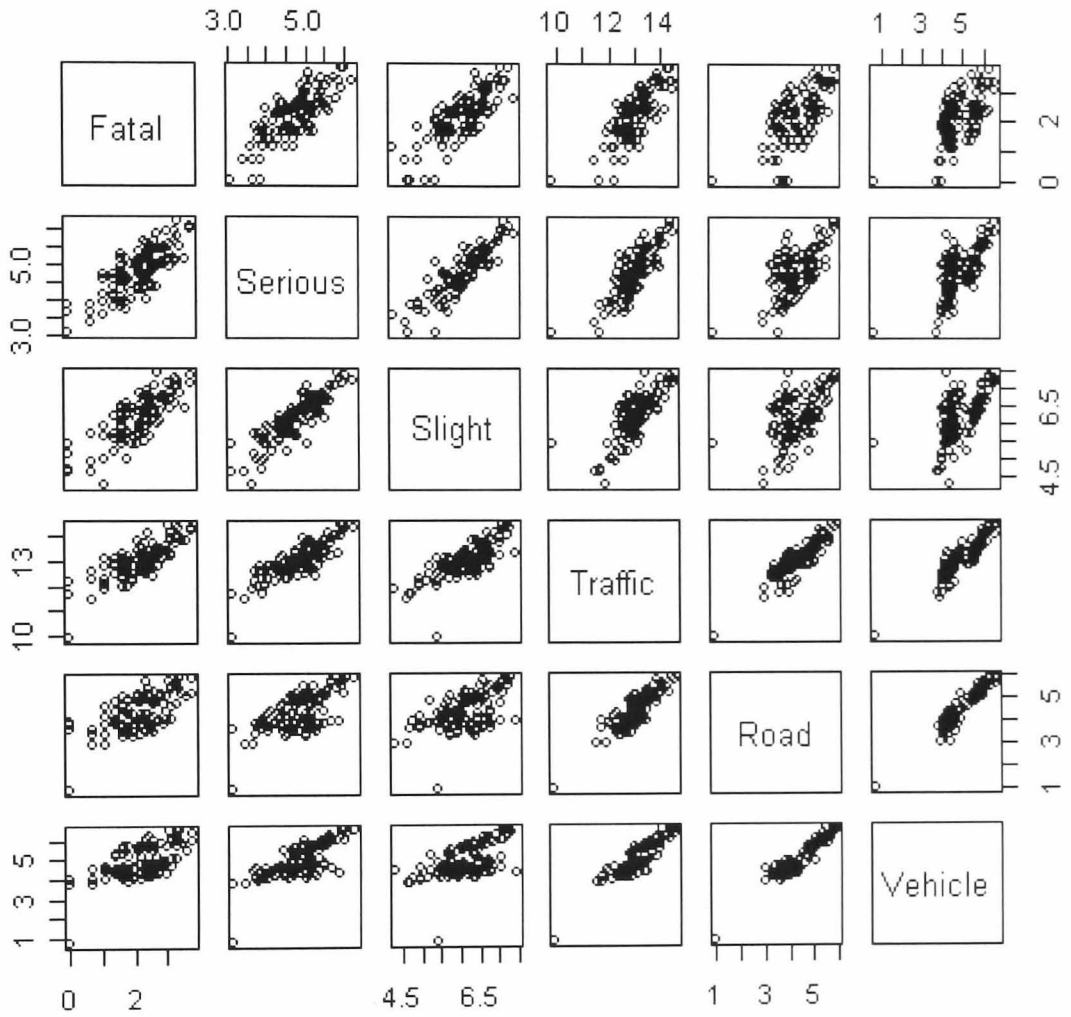

Figure 6.1: Relationships between the variables in logarithmic forms for accidents in 1986: 'Fatal' for fatal accidents; 'Serious' for serious accidents; 'Slight' for slight accidents; 'Traffic' for traffic volume in million vehicle-km; 'Road' for road length in km; 'Vehicle' for number of registered vehicles in thousand.

6.2.1 Models for accidents in a single year

In order to find out to what extent the conventional models for road accidents can be improved by including a CAR prior, before using four years' data to fit the models, a study based on only one year's data was made first. Here, models were developed for

fatal and serious accidents separately. Tables 6.1 to 6.3 show some of the results from the models.

Table 6.1: Summary of the model fits for fatal accidents in 1986

Model	Metropolitan effects	DIC	Moran's I	
			value	p-value
PL	without	621.3	0.10	0.07
PLre	with	616.9	0.03	0.56
PLN	without	617.9	0.01	0.12
CCAR	without	685.0	-0.04	0.59

As explained in Section 4.6, a model with a lower DIC performs better. Table 6.1 shows that, for fatal accidents, the Poisson log-linear model with metropolitan county effects (PLre) and the Poisson log-linear model with log-normal random effects (PLN) have similar DICs. Their DICs are not very different from that of the Poisson log-linear model (PL). The DIC of the CAR model is the highest. These indicate that a CAR model is not appropriate here. In order to examine the spatial correlation in the residuals, Moran's I statistic was obtained for each model. A p-value less than 0.05 indicates that the spatial correlation in the residuals is significant. According to the table, all the p-values are larger than 0.05. Therefore, the spatial correlation in the residuals for fatal accidents is not significant. This result can be interpreted in two ways. First, it suggests that fatal accidents at the local authority level may not tend to be spatially dependent throughout the whole geographical area (for instance, a country). Secondly, the non-significant correlation may be just because numbers of fatal accidents at the local authority level in one year are too few. Therefore, using accident data in a longer period may lead to a different result. When using the total number of fatal accidents in 4 years as the response variable, the residuals from the models without considering any spatial effect were found to be spatially correlated. The spatial correlation can be removed by using a CAR model that also improves the DIC. This indicates that the nonsignificant spatial correlation in the residuals from the earlier models for fatal accidents in only one year is just due to the reason of sparse data.

Table 6.2 summarizes the results of models for serious accidents without a CAR component. The inclusion of metropolitan effects (see model PLre) greatly improves the DIC of the Poisson log-linear model (PL) and causes Moran's I to drop. For the Poisson-

Table 6.2: Summary of the model fits for serious accidents in 1986, excluding CAR models

Model	Metropolitan effects	DIC	Moran's I	
			value	p-value
PL	without	2097.3	0.25	0.00
PLre	with	1800.3	0.13	0.02
PLN	without	935.1	0.29	0.00
PLNre	with	929.3	0.14	0.01

regression model with log-normal random effects (PLN), this inclusion does not improve the DIC much(see model PLNre), but does cause Moran's I to drop although I is still significant and positive. The positive Moran's I indicates that there is positive spatial autocorrelation in the residuals and suggests that the spatial correlation in the accident means across local authorities is not completely explained by the explanatory variables.

Table 6.3: Summary of the model fits for CAR models for serious accidents in 1986

Model	definition of neighbour	length of neighbours list	choice of weights	DIC	Moran's I	
					value	p-value
$ICAR_{nb1}$	common boundary	550	1–0	934.6	-0.19	0.00
$CCAR_{nb1}$	common boundary	550	1–0	931.2	-0.15	0.01
$ICAR_{nb1dist}$	common boundary	550	Euclidean distance	944.3	-0.74	0.00
$CCAR_{nb1dist}$	common boundary	550	Euclidean distance	934.2	-0.21	0.00
$ICAR_{nb2}$	motorways and trunk roads	238	1–0	990.7	-0.09	0.36
$CCAR_{nb2}$	motorways and trunk roads	238	1–0	934.6	-0.16	0.05

Table 6.3 summarizes the results of some CAR models. For spatial models that use a CAR prior, models with names beginning with 'ICAR' correspond to the intrinsic CAR models that do not include the unstructured random effects in the models and those with names beginning with 'CCAR' correspond to convolution CAR models that include both the spatially structured random effects and the unstructured random effects. The subscript in the model names describes how the neighbours list is defined and what weighting scheme is adopted. 'nb1' means that the neighbours are defined as local authorities that share at least one common boundary. 'nb2' means that the neighbours list is defined by using the spatial layout of the motorways in England and the trunk roads in metropoli-

tan districts. 'nb3' means the neighbours are identified based on the road network of the trunk roads in England. 'dist' means the spatial weights are computed based on the Euclidean distance between the centroids of the neighbouring local authorities; otherwise, 1-0 weights are used. Detailed information about the choices for the neighbours list and spatial weights has been introduced in Section 5.3.1.

The 'length of neighbours list' is the sum of the numbers of neighbours for each local authority. The lengths of the neighbours lists for models ICAR_{nb2} and CCAR_{nb2} are the shortest. The length of the boundary-based neighbours list is 550 while the length of the neighbours list determined by the road network is 238. Their neighbours lists are determined by the layout of the motorways in England and the trunk roads in metropolitan districts.

According to the values of the DICs, none of the models in Table 6.3 performs better than model PLNre in Table 6.2. According to the Moran's I statistic, only model ICAR_{nb2} shows a nonsignificant result. Moran's I for other models is still significant, but turns negative. A possible reason is that too many spatial effects were introduced. As is shown in Table 6.3, the absolute values of Moran's I in models CCAR_{nb1} and $\text{CCAR}_{nb1dist}$ drop when including the unstructured random effects compared to models ICAR_{nb1} and $\text{ICAR}_{nb1dist}$. This indicates that adding unstructured random effects to an intrinsic CAR model provides a compromise scheme for models of spatially correlated data especially when a high degree of spatial dependent structure is used. For the intrinsic CAR models ICAR_{nb1} and ICAR_{nb2}, values of Moran's I are -0.19 and -0.09 respectively. This may indicate that using a neighbouring structure that defines more neighbours could introduce more spatial effects and sometimes could even over-introduce the spatial effects. Therefore, an appropriate definition of the neighbours list for a CAR model is very important. Results for models ICAR_{nb2} and CCAR_{nb2} in Table 6.3 show some positive evidence that the spatial layout of the road network is useful to identify the spatial dependent relationships in the accident means.

Residual maps in Figure 6.2 plot the standardized residuals from models PLNre, CCAR_{nb1}, and CCAR_{nb2} respectively. Details for London boroughs are shown in Fig-

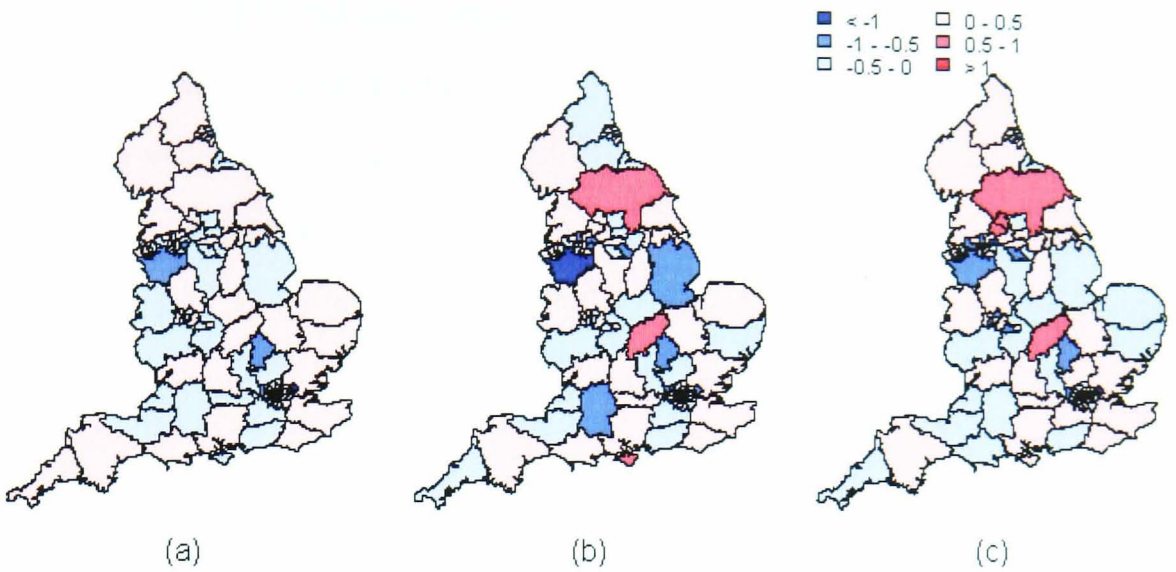

Figure 6.2: Maps (England) of standardized residuals for serious accidents: (a) model PLNre (Poisson model with log-normal random effects and metropolitan effects); (b) model CCAR_{nb1} (convolution CAR model whose neighbours list is determined by the boundaries); (c) model CCAR_{nb2} (convolution CAR model whose neighbours list depends on the layout of the road network).

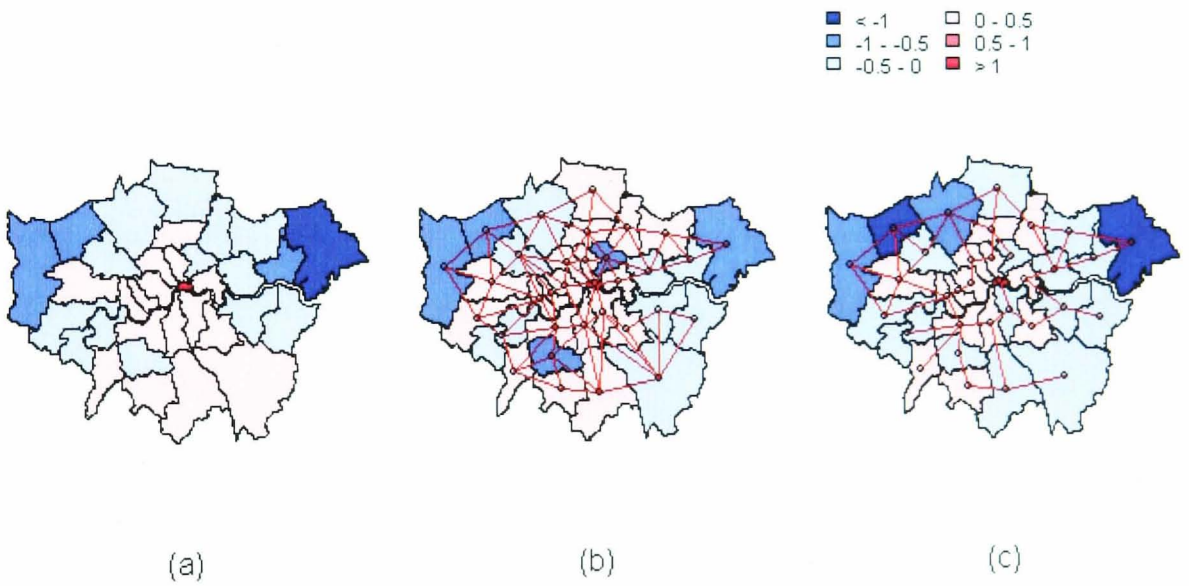

Figure 6.3: Maps (London boroughs) of standardized residuals for serious accidents – using the same models as in Figure 6.2.

ure 6.3. In order to distinguish the positive and negative residuals clearly, the residual plots adopt the diverging palettes based on the Hue-Chroma-Luminance (HCL) colour scheme suggested by Zeileis and Hornik (2006). This can be implemented by using the package 'vcd' in R (Meyer et al., 2007). In the last two maps in Figure 6.3, nodes in the maps correspond to the centroids of the areas. Links, which connect nodes, illustrate the

structure of the neighbours list. Local authorities that are connected by a direct link are defined as neighbours. These maps give a quick view of the changes in the concentrations of residuals with similar values when different models are used. Thus, they exhibit the progress in modelling the spatial correlation. The last two maps in both Figures 6.2 and Figure 6.3 indicate that the inclusion of a CAR component leads to a more random pattern in the residual map. The lack of apparent clustering, compared with the first map in both Figures 6.2 and Figure 6.3, indicates that CAR models perform successfully to account for the existing spatial correlation in the accident means.

The result from models using only one year's data shows that the inclusion of a CAR prior in the model for serious accidents removes the positive spatial correlation in the residuals from the models that do not consider the spatial random effects. Therefore, data in other years were used together to fit similar forms of models.

6.2.2 Models for four years' data

6.2.2.1 DIC and spatial correlation

Models were developed for accidents of different severity separately. Tables 6.4, 6.5 and 6.6 show some measures of the model performance and the spatial correlation in the residuals for some selected models.

For fatal accidents, Moran's I test is significant for residuals from model PL and PLfe in 1984 and 1986. When the metropolitan effects are included in the models, Moran's I for these two years become nonsignificant. The DICs are gradually improved by adding additional fixed or random effects in the Poisson log-linear model, which is shown in Table 6.4. However, model PLNtr, which includes a linear time trend variable, has a poor DIC. Its expected deviance and effective number of parameters are the highest among all the models. Using a CAR prior to take account of spatial random effects does not improve the DIC compared with model PLNre, which corresponds to the Poisson log-linear model with metropolitan county effects. The best performing model is PLNre&temp2, which extends model PLNre by including the temporal effects modelled by a first order autoregressive prior (see model (4.4) in Section 4.3). Using a random walk, formulated by a

CAR prior (see Section 4.3), to model the temporal effects does not improve the DIC.

Table 6.4: Summary of the model fits for fatal accidents (1983-1986)

Model	Length of neighbours list	DIC	Expected deviance	Effective number of parameters	Moran's I 1983	1984	1985	1986
PL		2482	2476	6	0.03	0.12(*)	0.01	0.08(*)
PLfe		2462	2454	8	0.02	0.11(*)	−0.02	0.08(*)
PLre		2436	2423	13	0.00	0.03	−0.03	0.02
PLN		2394	2242	152	0.00	0.11(*)	−0.01	0.07
PLNre		2375	2240	135	0.00	0.03	−0.04	0.01
PLNtr		2724	2240	484	0.00	0.11(*)	−0.01	0.07
$CCAR_{nb1}$	550	2392	2243	149	0.00	0.04	0.00	0.05
$CCAR_{nb3road}$	452	2386	2239	147	0.03	0.02	−0.02	0.02
PLNre&temp1		2377	2235	142	0.00	0.11(*)	−0.01	0.07
PLNre&temp2		2325	2213	112	0.00	0.10(*)	−0.01	0.06

*: significant at the 5% level

Table 6.5: Summary of the model fits for serious accidents (1983-1986)

Model	Length of neighbours list	DIC	Expected deviance	Effective number of parameters	Moran's I 1983	1984	1985	1986
PL		8687	8681	6	0.23(*)	0.23(*)	0.43(*)	0.44(*)
PLfe		7938	7930	8	0.28(*)	0.24(*)	0.37(*)	0.35(*)
PLre		7768	7755	13	0.22(*)	0.17(*)	0.33(*)	0.31(*)
PLN		3699	3308	391	0.21(*)	0.23(*)	0.40(*)	0.32(*)
PLNre		3693	3307	386	0.19(*)	0.18(*)	0.28(*)	0.21(*)
PLNtr		3732	3310	422	0.21(*)	0.24(*)	0.41(*)	0.32(*)
$ICAR_{nb1}$	550	3734	3364	370	−0.08	−0.07	−0.04	−0.11
$ICAR_{nb2}$	238	4021	3699	322	−0.08	−0.04	−0.06	−0.10
$ICAR_{nb3}$	452	3725	3356	369	−0.20(*)	−0.13	−0.11	−0.23(*)
$ICAR_{nb1dist}$	550	3749	3367	382	−0.32(*)	−0.29(*)	−0.53(*)	−0.42(*)
$ICAR_{nb3road}$	452	3732	3364	368	−0.24(*)	−0.20(*)	−0.15(*)	−0.28(*)
$CCAR_{nb1}$	550	3683	3305	378	0.03	0.01	0.07	−0.01
$CCAR_{nb2}$	238	3683	3307	376	0.04	0.04	0.08	0.00
$CCAR_{nb3}$	452	3679	3302	377	−0.07	−0.02	−0.02	−0.15
$CCAR_{nb1dist}$	550	3686	3297	389	−0.19(*)	−0.16(*)	−0.31(*)	−0.37(*)
$CCAR_{nb3road}$	452	3673	3298	375	−0.04	−0.03	0.00	−0.14(*)
$CCAR(t)_{nb3road}$	452	3614	3303	311	−0.28(*)	−0.13(*)	−0.34(*)	−0.16(*)
$CCAR(t)_{nb3road}$temp1	452	3673	3297	376	−0.05	−0.03	0.01	−0.15(*)
$CCAR(t)_{nb3road}$temp2	452	3617	3291	326	−0.04	−0.05	0.05	−0.14(*)

*: significant at the 5% level

For serious accidents, by adding fixed spatial effects to describe the local authority type (model PLfe) and metropolitan county effects (model PLre) in the Poisson log-linear model PL, the DICs decrease gradually (see Table 6.5). The inclusion of a time trend variable again gives a worse DIC. For all the models that do not include spatial random effects, the Moran's I is large and positive, especially in year 1985 and 1986. Results indicate that the amount of the spatial correlation in the residuals for serious accidents is higher than that for fatal accidents. The inclusion of the fixed spatial effects can make Moran's I drop for all the years, but Moran's I is still large and significant. The DIC is

much improved when log-normal random effects are included (see model PLN). None of the intrinsic CAR models that include only the spatially structured random effects perform better than model PLNre. At the same time, most of the values of Moran's I turn negative. The same problem has risen in the study using only one year's data in the previous subsection. The empirical work of Cliff and Ord (1981) shows that if the autoregressive model residuals are substantially the same as the OLS regression residuals, then they will probably turn out to be significantly negatively autocorrelated as measured by Moran's I. This may explain why a significant negative Moran's I is obtained after including a CAR component in the model in this study. Another possible reason of obtaining negative value of Moran's I is that too many random spatial effects were introduced. After taking account of both the spatially structured random effects and the unstructured random effects, the absolute values of Moran's I for most of the CCAR models in Table 6.5 are smaller. However, for some of these CCAR models, the values of Moran's I are still negative and large especially in year 1986. This is consistent with the earlier result using only data for serious accidents in 1986.

The DICs of all the CCAR models are close. In the last three CAR models, neighbours are defined based on the layout of the trunk roads in England (see Section 4.2.3). The best performing CCAR model is model CCAR(t)$_{nb3road}$ according to the DIC. The spatial weights for it are based on the number of the common roads that crossed the neighbouring local authorities. However, spatial correlation in the residuals measured by Moran's I from this model is negative and significant for all the years. This indicates that too many spatial effects might be introduced. Moreover, the model assumes that the variance of the spatial random effects varies in time. When a constant variance parameter was used during the study period, the convergence of several parameters is poor.

The last two CAR models in Table 6.5 include temporal effects. By using a random walk to model the temporal effects (model CCAR(t)$_{nb3road}$temp1), the DIC of the model is worse than that of model CCAR(t)$_{nb3road}$. Applying an autoregressive prior to model the temporal effects (model CCAR(t)$_{nb3road}$temp2) results in a similar DIC compared with that of model CCAR(t)$_{nb3road}$. By using this model, the spatial correlation in the residuals

for most years is nonsignificant. This indicates that, after taking account of the temporal autoregressive effects, more appropriate level of spatial effects was introduced.

Table 6.6: Summary of the model fits for slight accidents (1983-1986)

Model	Length of neighbours list	DIC	Expected deviance	Effective number of parameters	Moran's I 1983	1984	1985	1986
PL		24316	24310	6	0.38(*)	0.33(*)	0.28(*)	0.32(*)
PLfe		16598	16590	8	0.27(*)	0.23(*)	0.21(*)	0.24(*)
PLre		15678	15665	13	0.27(*)	0.23(*)	0.22(*)	0.24(*)
PLN		4303	3880	423	0.19(*)	0.22(*)	0.21(*)	0.21(*)
PLNre		4305	3876	429	0.20(*)	0.08(*)	0.11(*)	0.14(*)
PLNtr		4353	3880	473	0.17(*)	0.21(*)	0.21(*)	0.19(*)
CCAR(t)$_{nb1}$	550	4295	3879	416	−0.05	0.01	−0.03	−0.03
CCAR(t)$_{nb3road}$	452	4289	3878	411	−0.06	−0.04	−0.02	−0.08
CCAR(t)$_{nb3road}$temp1	452	4291	3879	412	−0.06	−0.04	−0.02	−0.08
CCAR(t)$_{nb3road}$temp2	452	4252	3867	385	−0.01	−0.03	−0.02	0.01

*: significant at the 5% level

As is shown in table 6.6, the influence of including fixed and random spatial effects to the Poisson log-linear models for slight accidents is similar to that for serious accidents. Moran's I is significant when spatial random effects are not included in the models. For the CCAR models, Moran's I is all negative and nonsignificant. The last model CCAR$_{nb3road}temp2$ with temporal effects, modelled by a first order autoregressive prior, performs best.

As introduced in Section 3.1.2, the variance of the spatially structured random effects (τ_θ) and the variance of the unstructured random effects (τ_ε) control the strength of their effects respectively. Table 6.7 shows the result of these variance and their ratios for selected CAR models. Different variance parameters are used for the spatial effects in different years. The results of their estimates show they are similar. Therefore, the mean of the variances is used in the table. Ratio of τ_θ and τ_ε will reflect the relative strength of the spatial random effects against the unstructured heterogeneity. All the ratios are smaller than the $\bar{w}$—the average number of neighbours for each local authority. This indicates that the unstructured heterogeneity dominates the spatially structured random effects.

Residual maps for selected models are shown in Figures 6.4 to 6.11. Details for London boroughs are plotted separately. These maps not only give a quick view of the improvement of the models by including extra fixed and random effects but also show the changes of concentrations of different values of residuals from models without and with

Table 6.7: Summary of the variance parameters in selected CAR models

Model	Severity	τ_ε	$\bar{\tau}_\theta$	$\bar{w}$	$\bar{\tau}_\theta/\tau_\varepsilon$
$CCAR_{nb1}$	Serious	0.16	0.40	5.1	2.57
$CCAR_{nb3road}$	Serious	0.17	0.44	9.5	2.64
$CCAR_{nb3road}$temp2	Serious	0.17	0.23	9.5	1.36
$CCAR_{nb1}$	Slight	0.18	0.43	5.1	2.36
$CCAR_{nb3road}$	Slight	0.13	0.53	9.5	4.17
$CCAR_{nb3road}$temp2	Slight	0.15	0.24	9.5	1.55

a CAR component. Thus, they exhibit the progress in modelling the spatial correlation.

For fatal accidents, by using models PL and PLNre, no apparent clusters of residuals with similar values are found in the England maps and the London maps (see Figures 6.4 to 6.7). This is consistent with the low value of Moran's I in the residuals for fatal accidents as shown in Table 6.4. When the metropolitan county effects and the unstructured random effects are included in the model, corresponding to model PLNre, some areas in the map are plotted in a lighter colour (see Figures 6.5 and 6.7). This indicates that better estimates are obtained and the performance of the models are improved. For serious accidents, Figures 6.8 and 6.9 show the residual maps for model PLNre. In some parts of England, some apparent clusters of the residuals of similar values are identified in particular years. For instance, in 1983, positive residuals are clustered in the middle part of England. The east part of this cluster remains in 1984. However, no apparent clusters are found in the same region in 1985 and 1986. In the maps of London boroughs, negative residuals have shown in most outter boroughs in 1983. Positive residuals are clustered in inner London and extend to the north-west and the south-east in 1984. This cluster expands in 1985. The spatial pattern of residuals in 1986 is similar to that in 1984. Figures 6.10 and 6.11 show the residual plots for model $CCAR(t)_{nb3road}$temp2. The cluster of positive residuals in the middle of England in 1983 and 1984, when using model PLNre, are more or less broken after adopting the CAR model as shown in the plots. In London boroughs, the residuals also exhibit a more random pattern.

For slight accidents, residuals from a CAR model again exhibit a more random pattern in a map than those from a non-CAR model. Overall, the inclusion of the spatial random effects in the models can lead to a more random pattern in the residual maps. The lack

of apparent clustering indicates that CAR models perform successfully to account for the existing spatial correlation in the accident means.

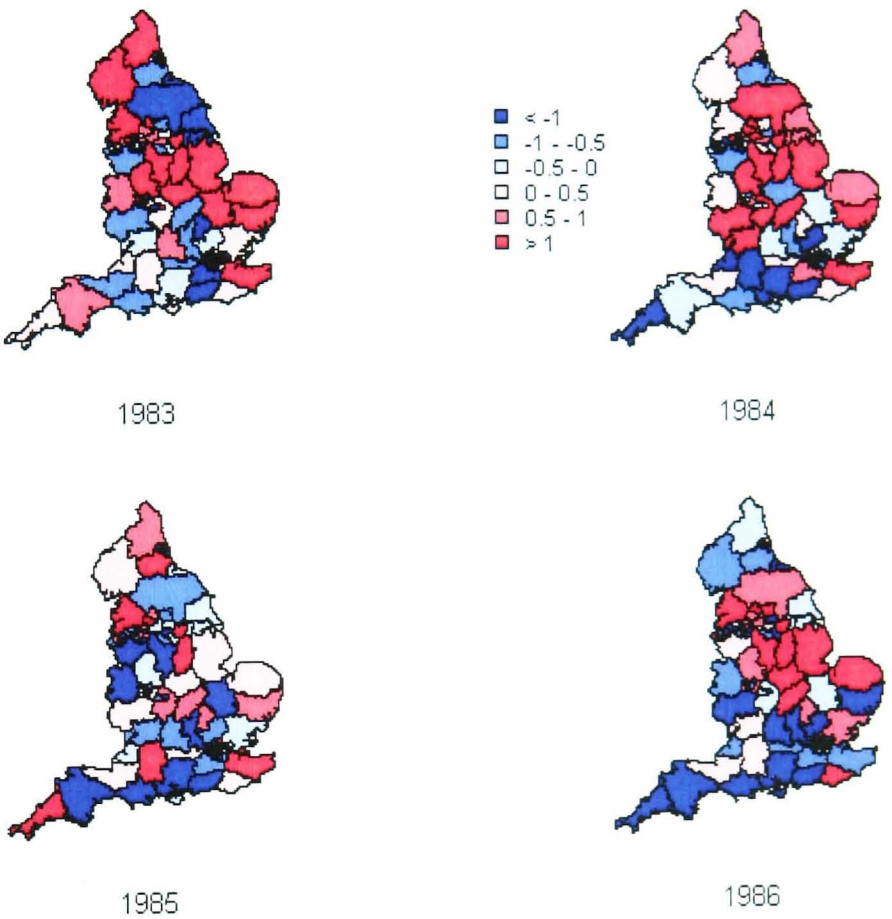

Figure 6.4: Residual maps for model PL (Poisson log-linear model): fatal accidents.

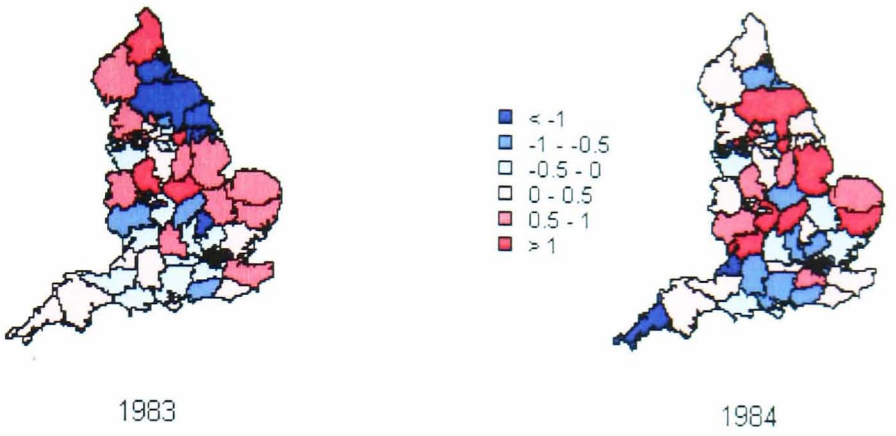

Figure 6.5: Residual maps for model PLNre (Poisson model with log-normal random effects and metropolitan county effects): fatal accidents.

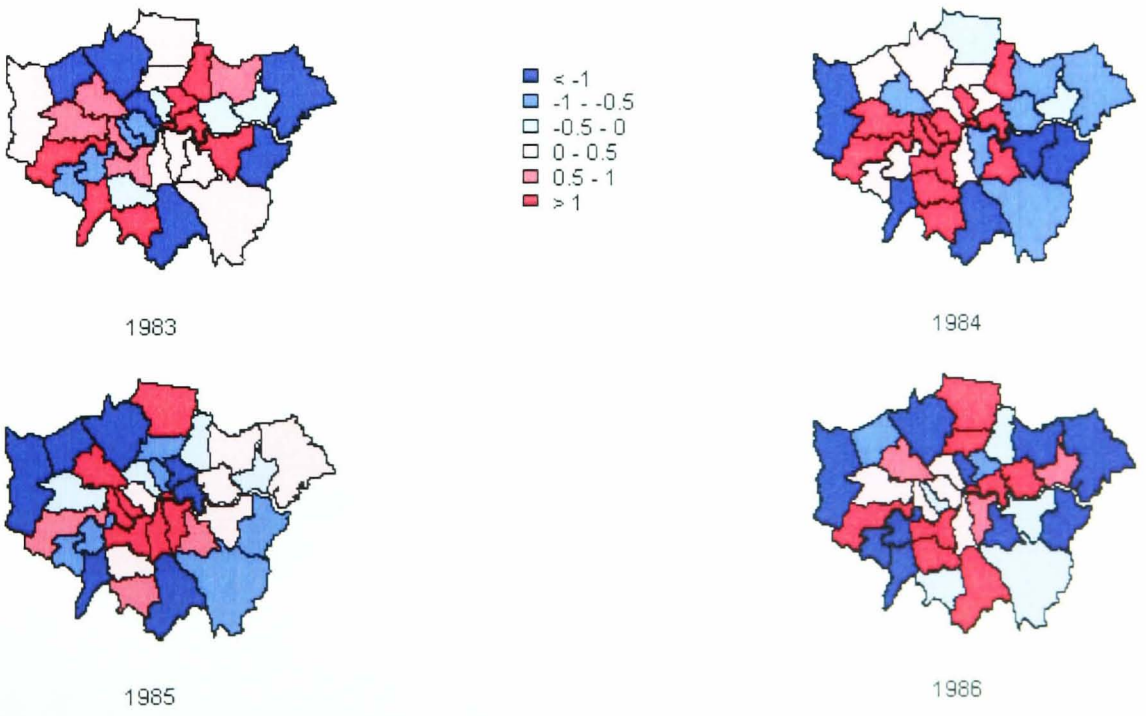

Figure 6.6: Residual maps of London boroughs for model PL: fatal accidents.

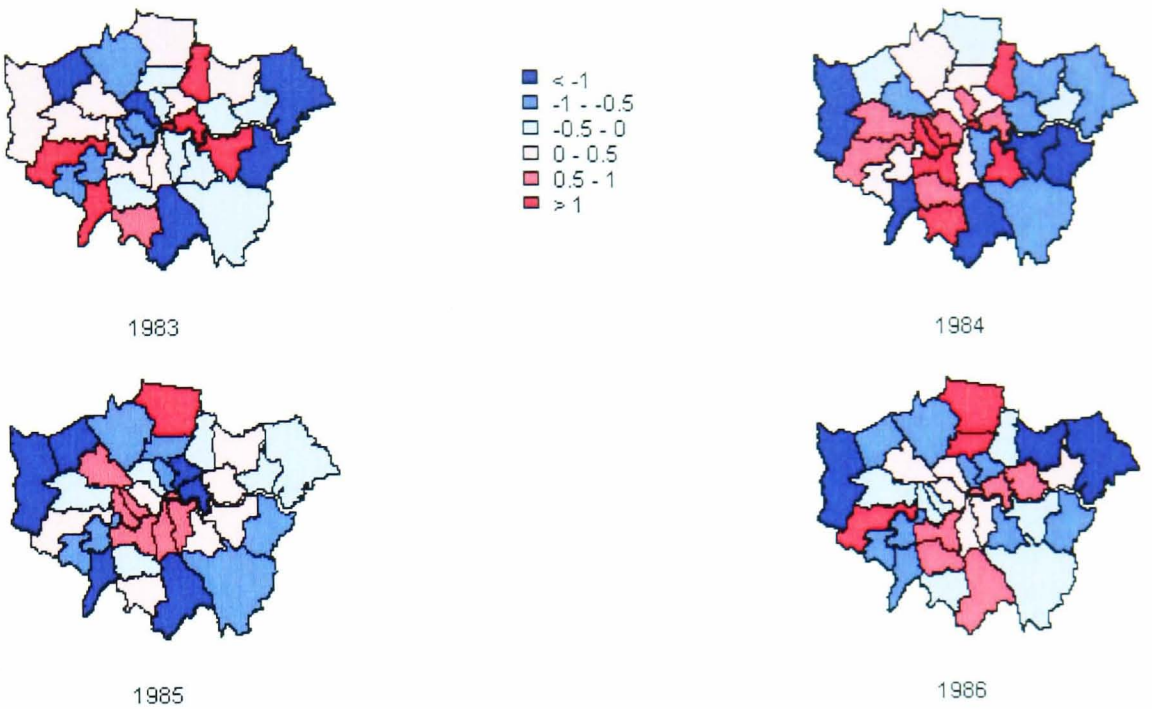

Figure 6.7: Residual maps of London boroughs for model PLNre: fatal accidents.

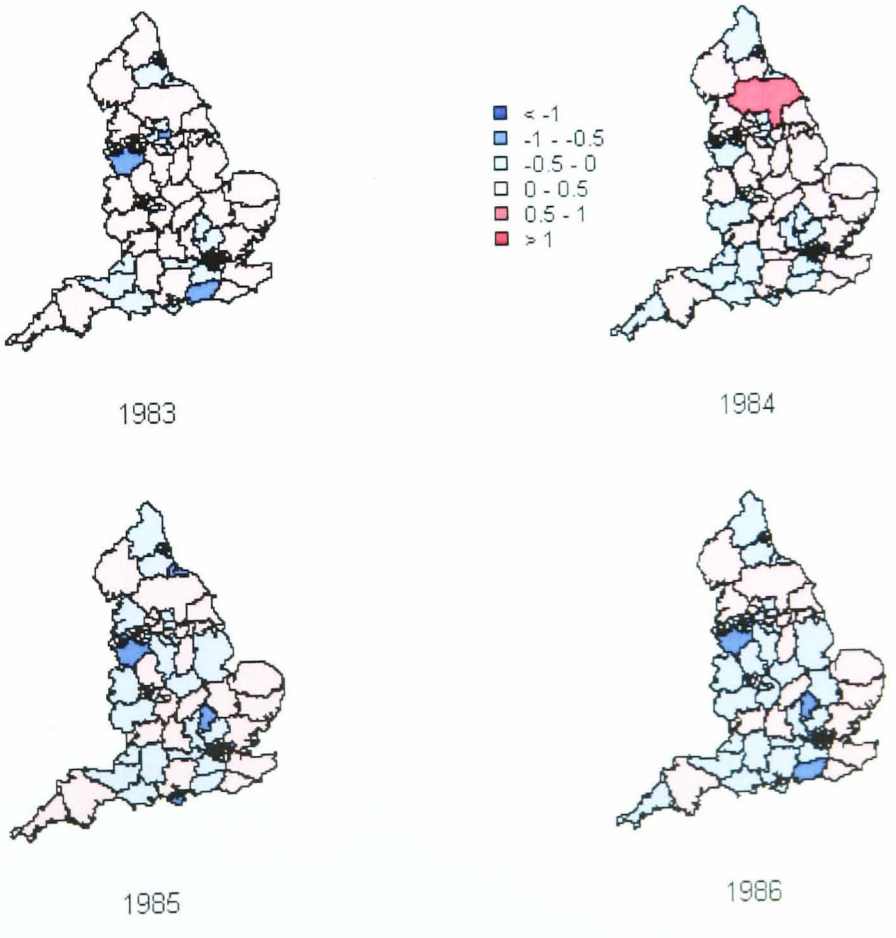

Figure 6.8: Residual maps for model PLNre (Poisson model with log-normal random effects and metropolitan county effects): serious accidents.

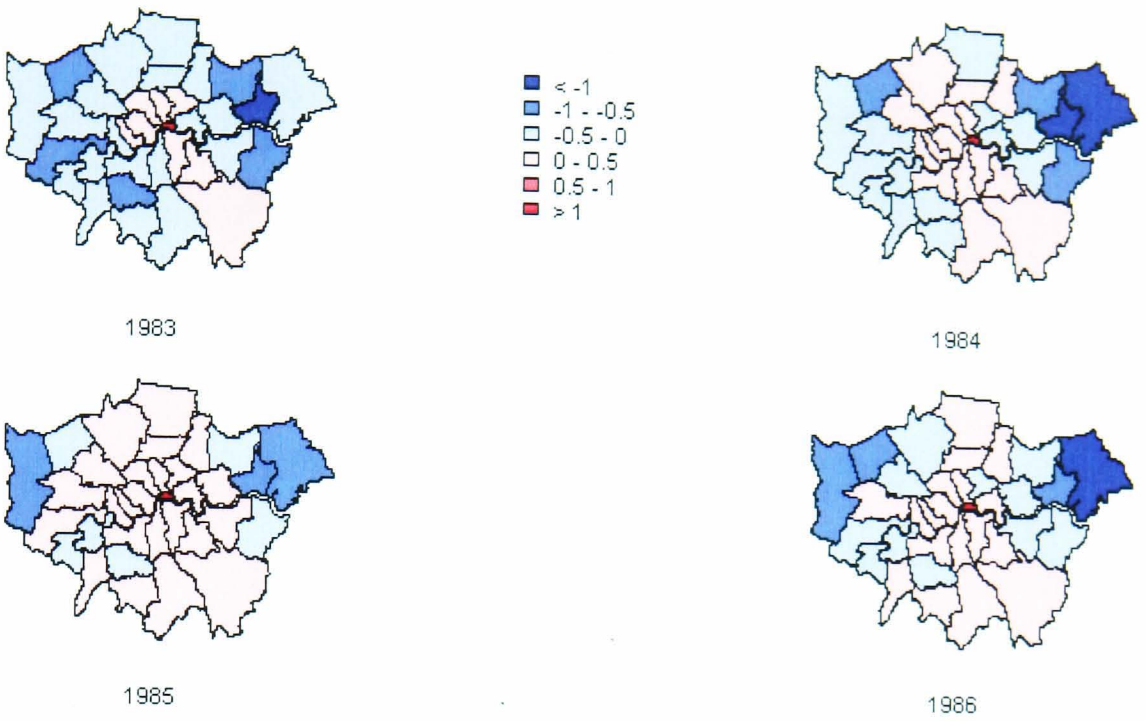

Figure 6.9: Residual maps of London boroughs for model PLNre: serious accidents.

6.2.2.2 Estimated parameters of the explanatory variables

Summaries of the estimated parameters for selected models can be found in tables in Appendix B. The tables include the estimated coefficients for the explanatory variables, namely area, population, number of licensed vehicles, road length and traffic respectively. As explained in Chapter 3, for Bayesian models, the estimates of the parameters can be given by their posterior medians and 2.5% and 97.5% percentiles. $\hat{R}$, measuring the convergence status for each parameter, is also included in the tables. Within 20,000 to 40,000 iterations, according to the values of $\hat{R}$, all the parameters of interest have satisfactorily converged. As shown earlier, the sign of the coefficients for the explanatory variables tells whether the relationships between the response variables and the explanatory variables are positive or negative. According to the scatter plots in Figure 6.1, the explanatory variables are positively correlated with the response variables. However not all the estimated medians of the coefficients are positive. Figures 6.12 to 6.18 show the 95% credible intervals of the coefficients for selected models. The signs of the posterior medians and whether the credible intervals include zero are clearly shown in these figures.

For fatal accidents, as shown in Table 6.4, model PLNre&temp2 performs best. It is a

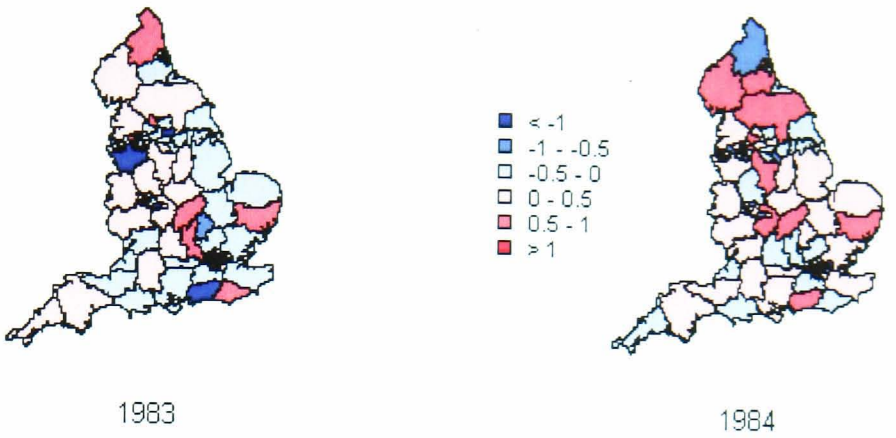

Figure 6.10: Residual maps for model CCAR(t)$_{nb3road}temp2$ (convolution CAR model with temporal effects, modelled by a first order autoregressive prior and its neighbours list depends on the layout of the road network): serious accidents.

Poisson model with log-normal random effects, metropolitan county effects and temporal effects modelled by a first order autoregressive prior. Figure 6.12 shows that none of the intervals for the coefficients of the explanatory variables includes zero except for road length. The medians for population, road length and traffic volume are all positive while those for area and number of licensed vehicles are negative. The negative coefficient for area can be explained as if the total road length is fixed, then the road density is less when the area is larger. Therefore, there could be less accidents. The coefficient for number of licensed vehicles is negative and has large variation. This could be due to the high correlation between it and other explanatory variables. The intervals of the coefficients for the metropolitan effects (see Figure 6.13) show that most intervals contain zero. For Merseyside, the interval include only negative values and shows large variation. These may indicate that the metropolitan county effects do not contribute much to explain the

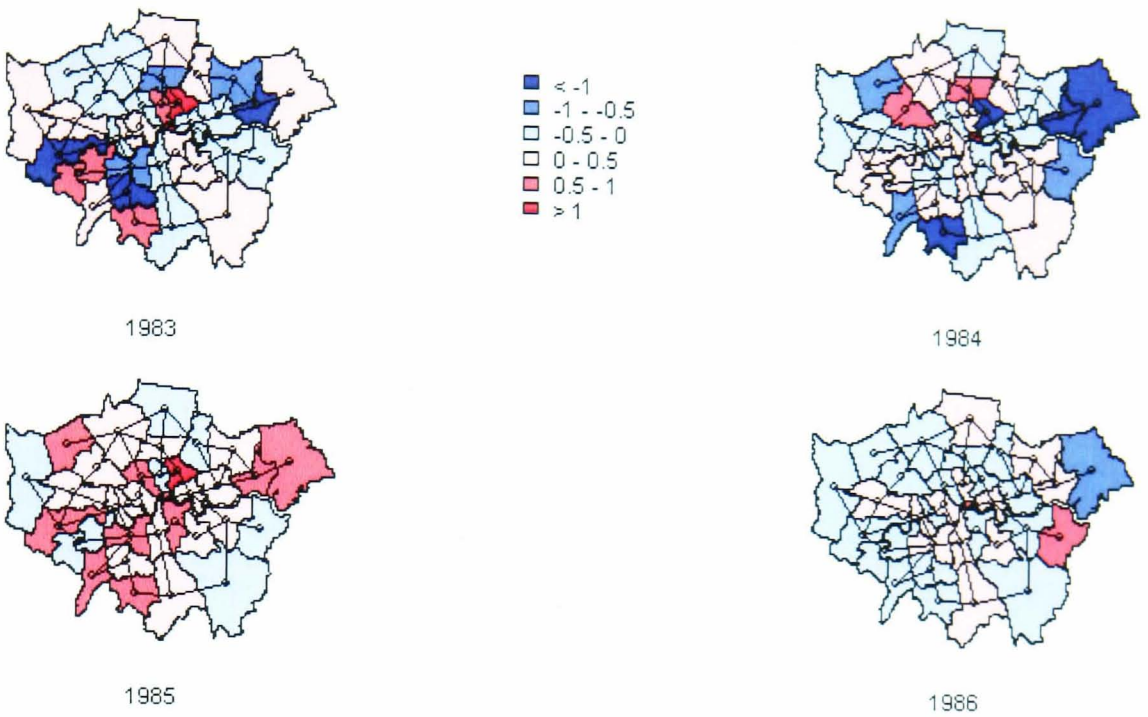

Figure 6.11: Residual maps of London boroughs for model CCAR(t)$_{nb3road}temp2$: serious accidents.

variation in the response variable. Therefore, no particular districts in a metropolitan county are likely to be associated with larger numbers of accidents.

For serious accidents, Figures 6.14 and 6.15 show the credible intervals of the coefficients for model PLN and one of the best performing models CCAR(t)$_{nb3road}$temp2. Model PLN includes log-normal random effects. Model CCAR(t)$_{nb3road}$temp2 is a convolution CAR model whose neighbours list depends on the layout of the road network. It also includes temporal effects modelled by a first order autoregressive prior. The credible intervals in the two figures look similar except for the variable population. The interval of the coefficient for area contains only negative values while those for number of licensed vehicles, road length and traffic volume contain only positive values. The reason for obtaining negative coefficient for area has been explained earlier. The interval of the coefficient for population in model PLN contains only negative values. In model CCAR(t)$_{nb3road}$temp2, it contains both positive and negative values with its median close to zero, indicating that the variable population does little to explain the variation in the response variable. When the metropolitan effects are included model PLN, all the intervals of the coefficients for the metropolitan effects contain zero (see Figure 6.16). This may

suggest that no particular metropolitan counties are associated with high or low accident frequencies.

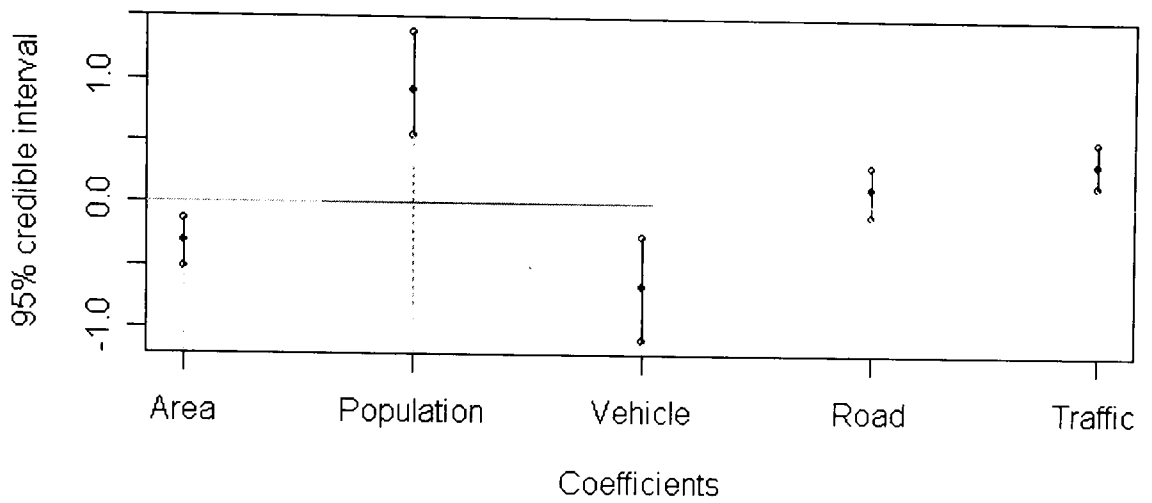

Figure 6.12: 95% credible intervals of the coefficients for the explanatory variables in model PLNre&temp2 for fatal accidents.

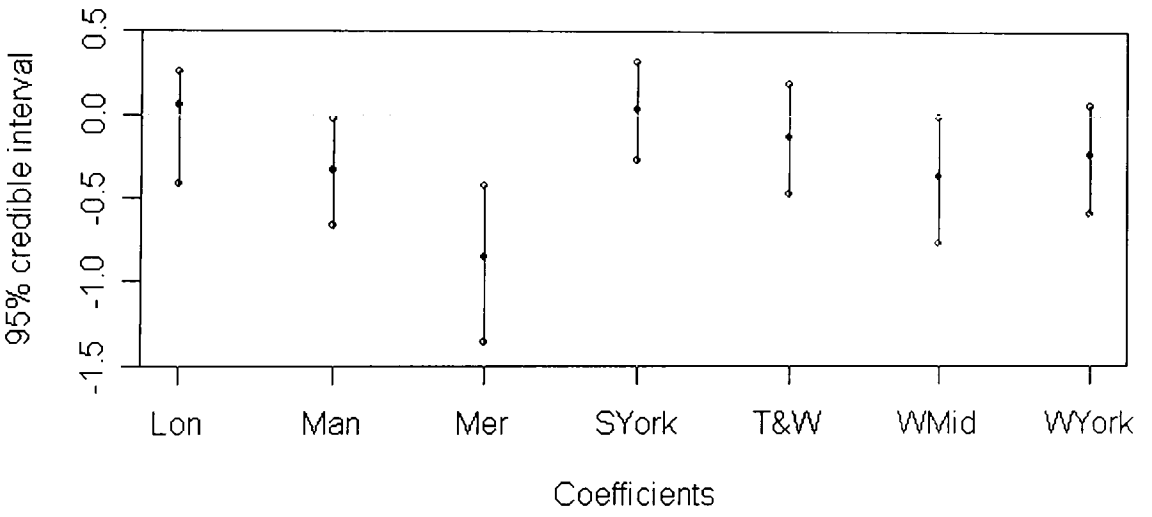

Figure 6.13: 95% credible intervals of the coefficients for dummy variables in model PLNre&temp2 for fatal accidents: 'Lon' for London boroughs; 'Man' for Great Manchester; 'Mer' for Merseyside; 'SYork' for South Yorkshire'; 'T&W' for Tyne and Wear; 'WMid' for West Midlands; 'WYork' for West Yorkshire.

Figures 6.17 and 6.18 show the credible intervals for slight accidents adopting the same models used for serious accidents. The interval of the coefficient for traffic only

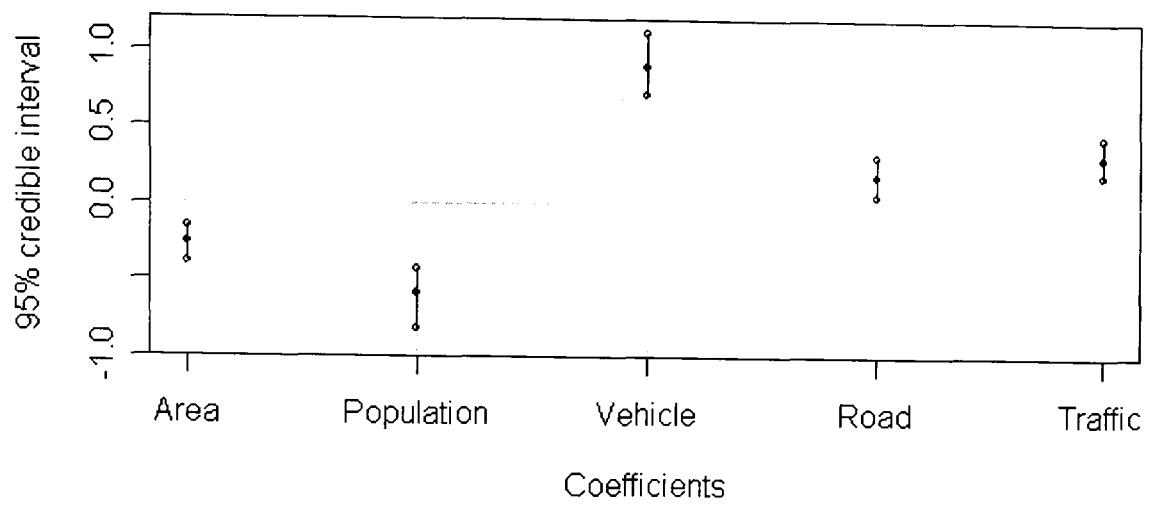

Figure 6.14: 95% credible intervals of the coefficients for the explanatory variables in model PLN: serious accidents.

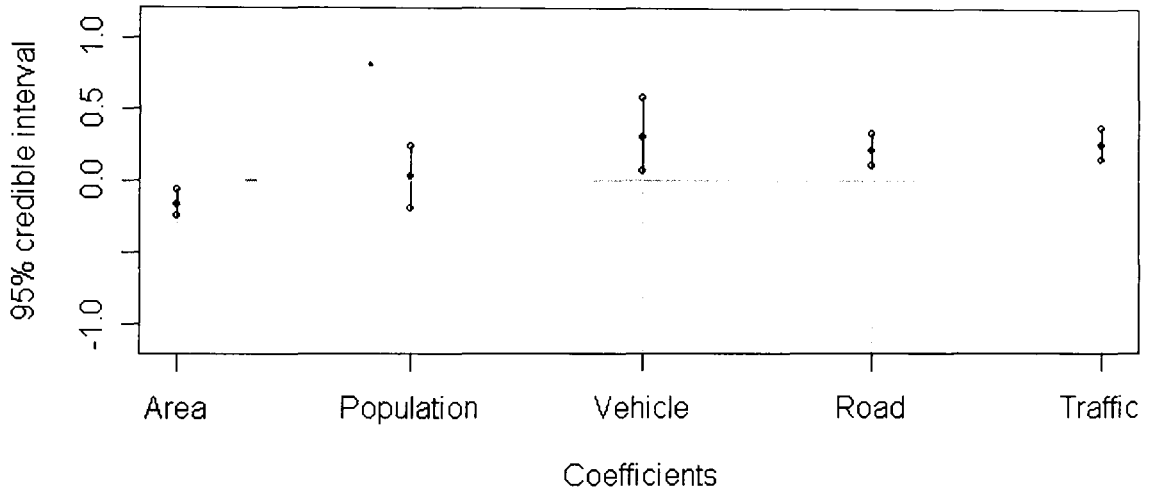

Figure 6.15: 95% credible intervals of the coefficients for the explanatory variables in model CCAR(t)$_{nb3road}$temp2: serious accidents.

contains positive values for both models. For other variables except for area in the CAR model, their coefficients either have large variation in their estimates or have both positive and negative values in their credible intervals, suggesting that they do not contribute much to explain the variation in the response variable.

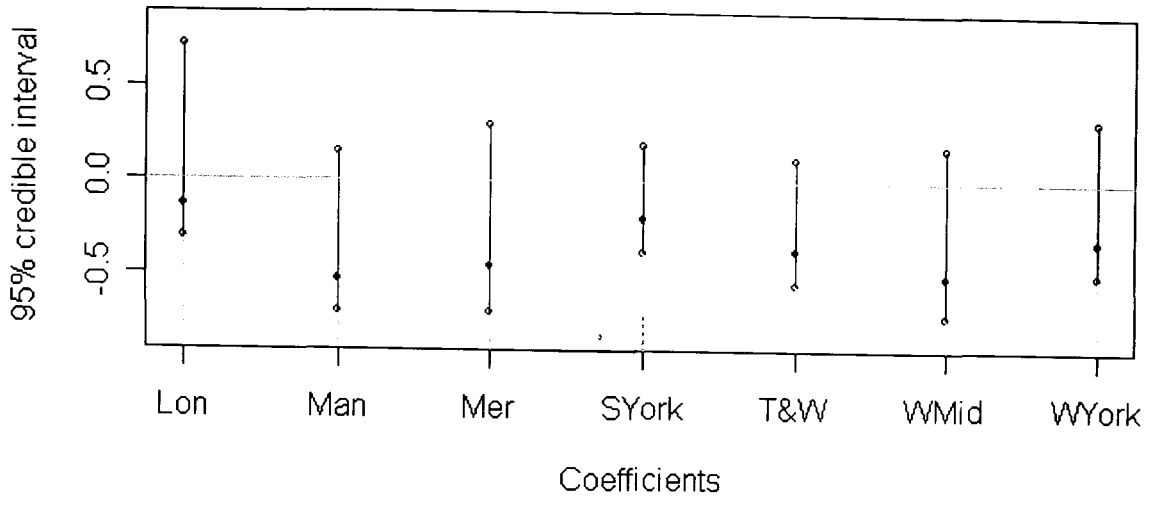

Figure 6.16: 95% credible intervals of the coefficients for the dummy variables in model PLNre: serious accidents (for full names of metropolitan counties, see Figure 6.13).

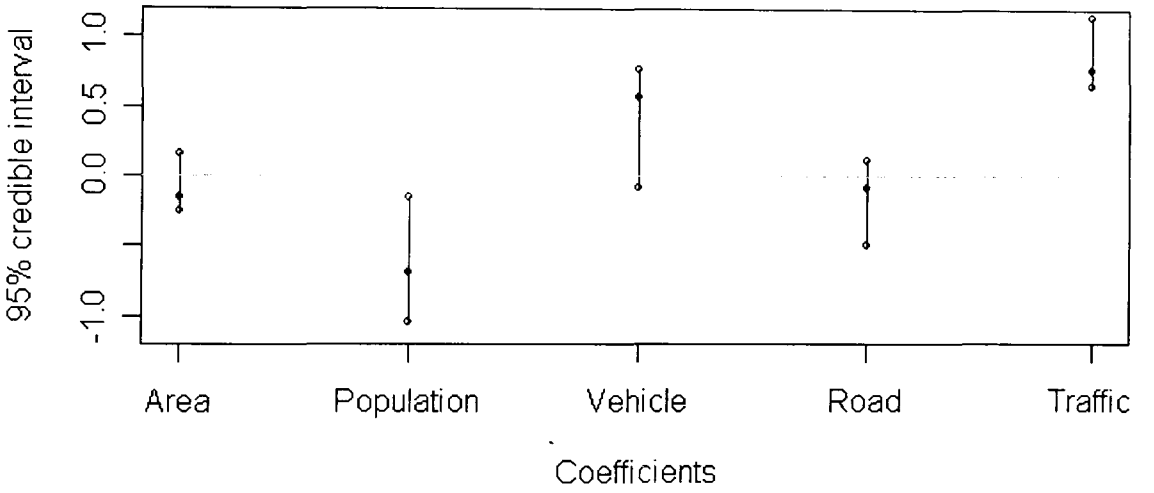

Figure 6.17: 95% credible intervals of the coefficients for the explanatory variables in model PLN: slight accidents.

6.2.2.3 Temporal correlation

In order to examine the use of including temporal effects, Pearson correlation coefficients were calculated for the four years' residuals from the selected models. They are shown in Tables 6.8 to 6.12

When no temporal effects are included in the models, the temporal correlation in the

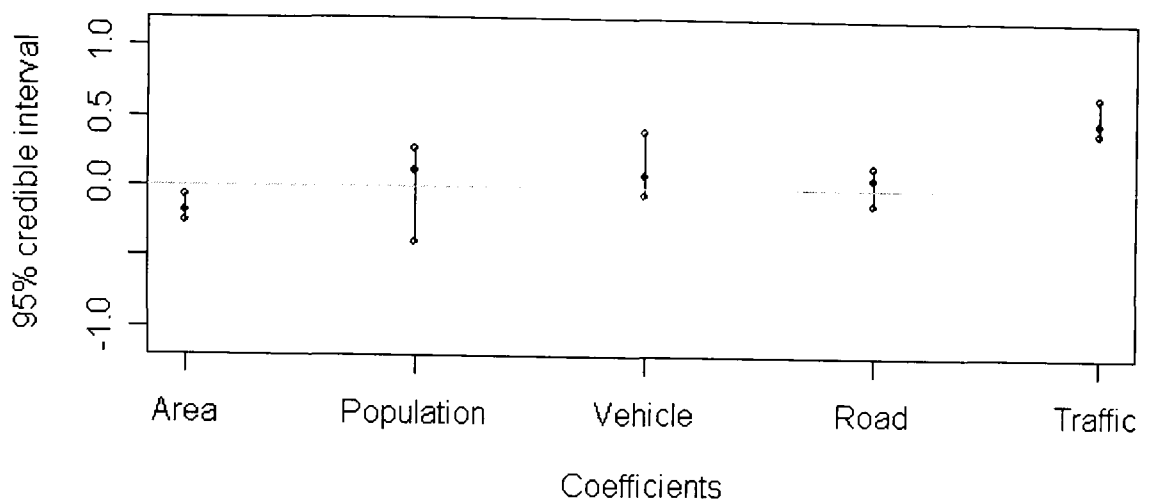

Figure 6.18: 95% credible intervals of the coefficients for the explanatory variables in model CCAR(t)$_{nb3road}$temp2: slight accidents.

residuals is positive for all types of accidents. Table 6.8 shows that the correlation coefficients of temporal correlation are not high for the fatal accidents. For serious accidents, the correlation matrix shown in Table 6.9 indicates there is large temporal correlation in the residuals. The temporal correlation in the residuals for slight accidents shown in 6.10 is even larger. This indicates that the inclusion of temporal effects to take account of the spatial correlation is necessary for serious and slight accidents. Tables 6.11 and 6.12 show the correlation matrix in the residuals from models that include both the spatial random effects and the temporal effects modelled by a first order autoregressive prior. The values of the correlation coefficients become smaller but some of them are negative. This may indicate that the temporal correlation is over-introduced in the models. The posterior median of parameter ρ, which accounts for the temporal correlation (see model (4.4) in Section 4.3), is 0.98 and 0.99 for serious accidents and slight accidents respectively.

Table 6.8: Temporal correlation coefficients for residuals from model PLNre for fatal accidents

Year	1983	1984	1985	1986
1983	1	0.29	0.18	0.30
1984	0.29	1	0.26	0.35
1985	0.18	0.26	1	0.35
1986	0.30	0.35	0.35	1

Table 6.9: Temporal correlation coefficients for residuals from model CCAR(t)$_{nb3road}$ for serious accidents

Year	1983	1984	1985	1986
1983	1	0.76	0.66	0.60
1984	0.76	1	0.78	0.69
1985	0.66	0.78	1	0.80
1986	0.60	0.69	0.80	1

Table 6.10: Temporal correlation coefficients for residuals from model CCAR(t)$_{nb3road}$ for slight accidents

Year	1983	1984	1985	1986
1983	1	0.82	0.83	0.80
1984	0.82	1	0.83	0.84
1985	0.83	0.83	1	0.90
1986	0.80	0.84	0.90	1

Table 6.11: Temporal correlation coefficients in residuals for serious accidents from model CCAR(t)$_{nb3road}$temp2

Year	1983	1984	1985	1986
1983	1	0.04	−0.07	−0.05
1984	0.04	1	−0.35	−0.19
1985	−0.06	−0.35	1	−0.38
1986	−0.05	−0.19	−0.39	1

Table 6.12: Temporal correlation coefficients in residuals for slight accidents from model CCAR(t)$_{nb3road}$temp2

Year	1983	1984	1985	1986
1983	1	0.14	0.20	−0.08
1984	0.14	1	−0.20	0.05
1985	0.20	0.20	1	−0.70
1986	−0.08	0.05	−0.70	1

6.3 Models for accidents at the local authority level in England from 2001 to 2005

In the previous section, models were fitted using data in 1980s. There are two main findings for the effect of including a CAR prior to take account of the spatial dependency in the local authorities. Firstly, the DIC is improved. Secondly, the positive spatial autocorrelation in the residuals from the non-CAR models, observed from both the Moran's I statistic and the residual maps, is removed. These suggest that for areal models the inclusion of a CAR prior is important. In this section, some up-to-date data are used to fit similar forms of models. In addition, more complicated models are considered. They take account of the correlation in different types of accidents.

The two response variables are the number of fatal and serious accidents, and the number of slight accidents in a local authority in a year. Eight explanatory variables are included in the models. They are area, population, length of A-roads, length of B-roads, length of other roads, number of junctions, car traffic and traffic of other vehicles. How the data for these variables were obtained has been explained in Section 5.3.2. Figure 6.19 shows the relationship of the response variables and selected explanatory variables in 2001. All the variables are in logarithmic forms. The figure shows that the two response variables are positively correlated and both of them are positively correlated with the explanatory variables. In addition, there are also high positive correlations between the explanatory variables. For other years, the relationship of the variables looks similar.

The national casualty reduction strategy in 2000 (see Department for Transport, 2001) established a road safety target in ten years. In order to achieve the final target, a gradual reduction in the casualties in each local authority is expected. This is closely associated with the reduction in the total number of accidents over time. Figures 6.20 to 6.21 show the trend in fatal and serious accidents, and slight accidents at the local authority level from 2001 to 2005. The drop in fatal and serious accidents during the period is particularly big for some local authorities. Figure 6.22 shows the trend in these local authorities. A search on the relevant statistics and reports published in their council websites shows that

6.3 Models for accidents at the local authority level in England from 2001 to 2005

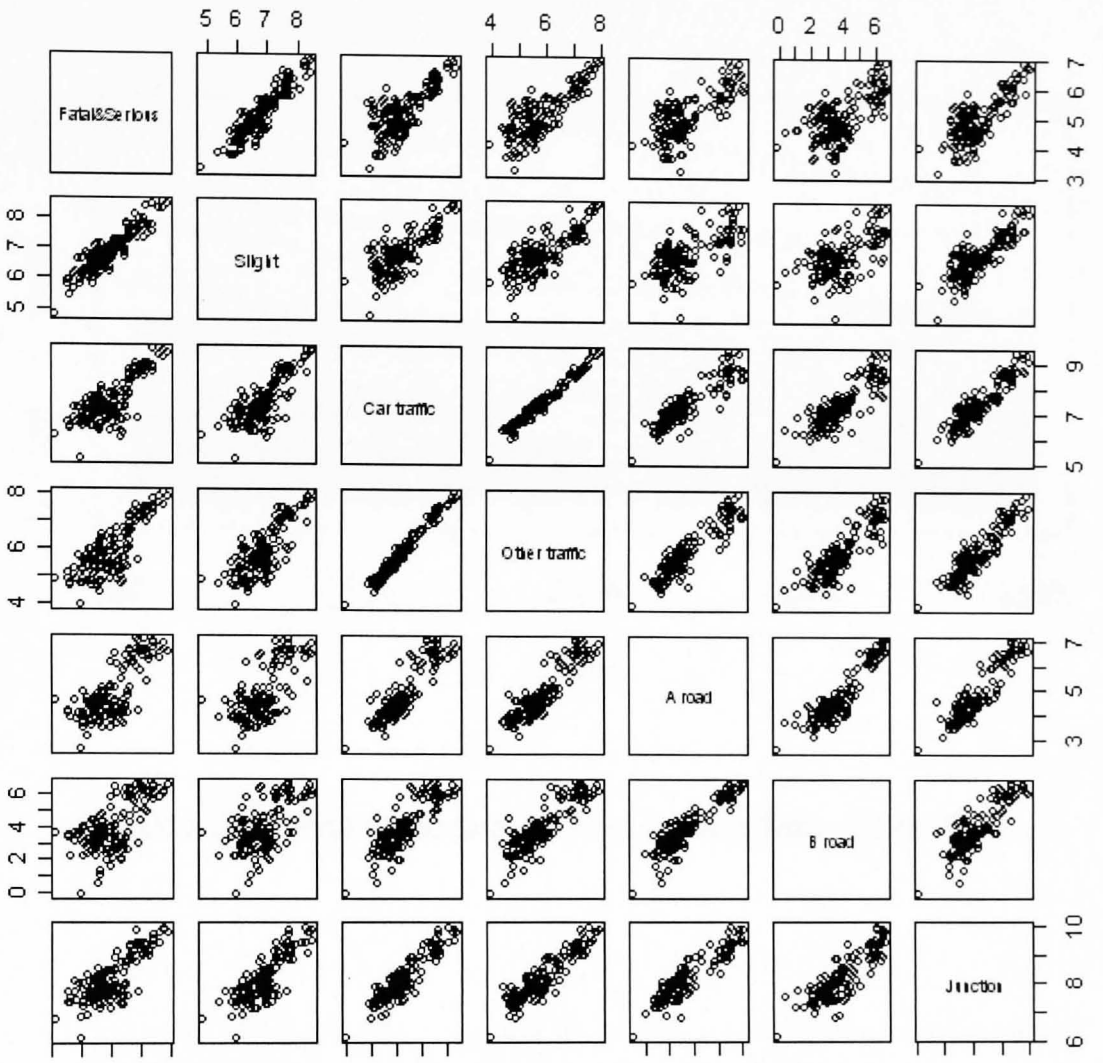

Figure 6.19: Relationship of the variables for fatal and serious accidents.

significant reduction in the accidents and casualties were claimed by the councils during the same period and the published statistics are consistent with the data obtained in this research.

6.3.1 Description of the models

Forms of the models for accidents at the local authority level in the 2000s are similar to the models used in the previous section. For models that do not include spatial random effects, a time trend variable and fixed spatial effects are added. For spatial CAR models, three types of models are examined. They are CAR models that have independent spatial random effects for accidents of different severity, CAR models that have spatial random

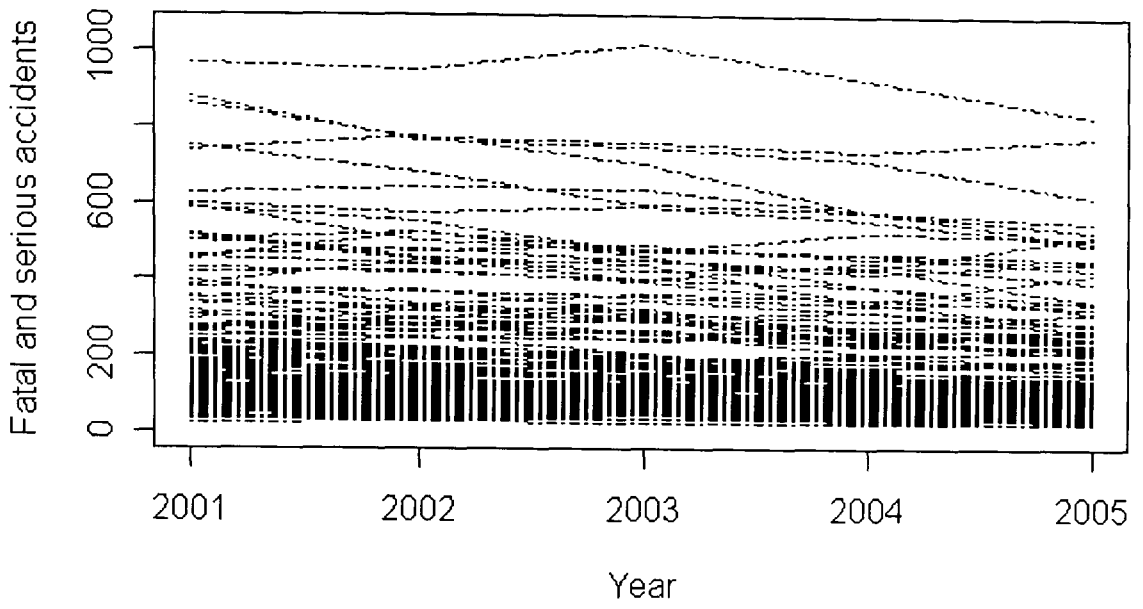

Figure 6.20: Trend in the fatal and serious accidents: 2001-2005.

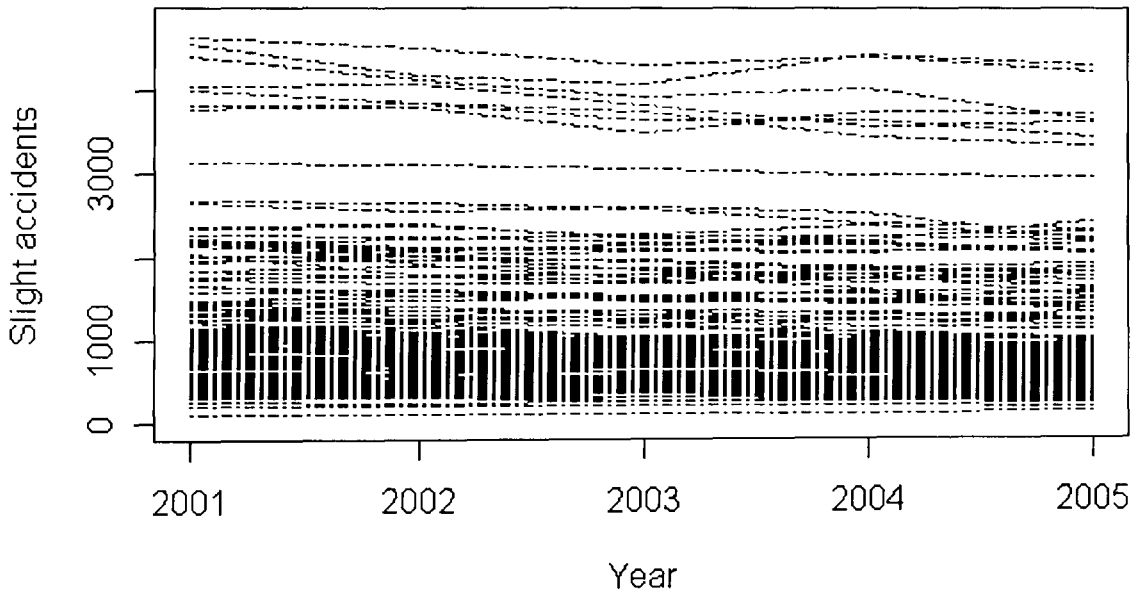

Figure 6.21: Trend in the slight accidents: 2001-2005.

effects which are correlated for different types of accidents, and spatial shared component CAR models. Moreover, temporal effects are considered and formulated with a first order autoregressive prior. Details of the forms of these models have been explained in

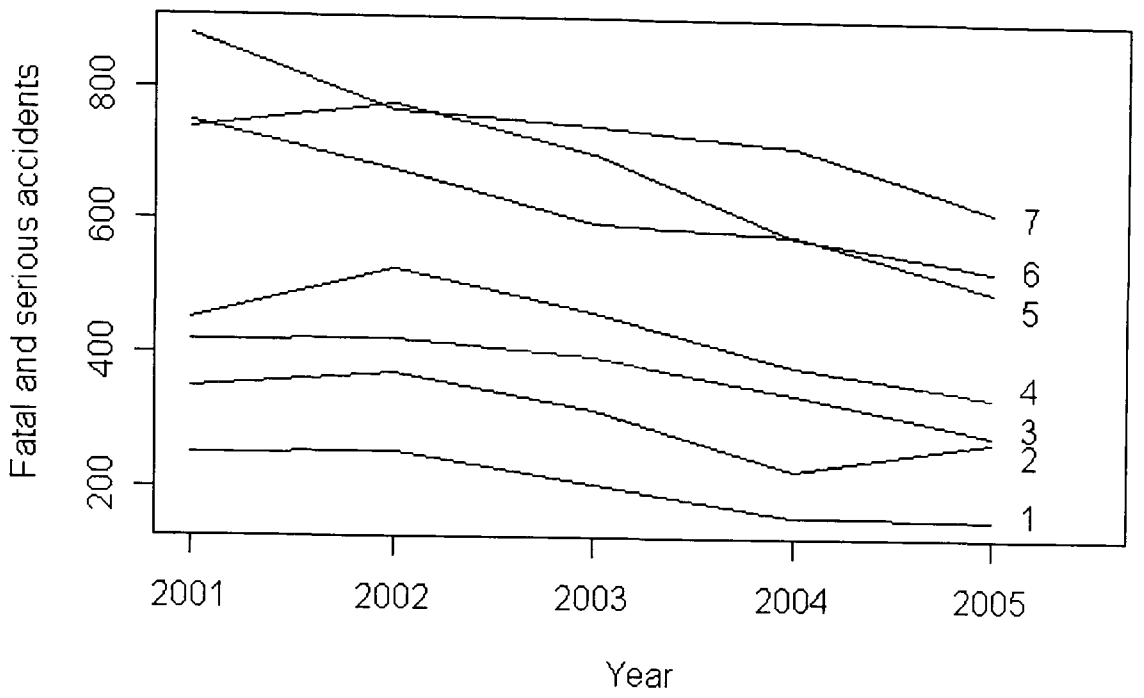

Figure 6.22: Trend in the fatal and serious accidents for selected local authorities: 1. Lambeth; 2. Devon; 3. Lincolnshire; 4. Oxfordshire; 5. Hertfordshire; 6. Hampshire; 7. Kent

Chapter 4. Only one neighbouring structure is used here. It defines neighbours based on the condition of sharing at least one common boundary. Only the 1-0 weighting scheme is considered.

6.3.2 DIC and spatial correlation

Table 6.13 summarises the model fits. The table shows that the DICs of the models

Table 6.13: Summary of the multivariate models for accidents in England in the 2000s

Model	DIC	Expected deviance	Effective number of parameters	Severity	*Moran's I* 2001	2002	2003	2004	2005
PLtr	41474	41454	20	fatal and serious	0.29(*)	0.22(*)	0.16(*)	0.02	0.05
				slight	0.29(*)	0.15(*)	0.13(*)	0.03	0.04
PLtr-re	38225	38189	36	fatal and serious	0.18(*)	0.12(*)	0.09(*)	0.03	0.12(*)
				slight	0.22(*)	0.09(*)	0.08(*)	0.02	0.05
PLNtr	19394	19072	322	fatal and serious	0.06	0.01	0.05	0.03	0.08
				slight	0.01	−0.03	0.04	0.04	0.10(*)
CCAR(t)tr	14424	13095	1329	fatal and serious	−0.06	−0.05	0.00	−0.09	0.00
				slight	0.03	−0.01	0.04	0.02	0.00
CCAR(t)tr.temp	14274	12930	1344	fatal and serious	−0.05	−0.04	0.00	−0.09	0.00
				slight	0.03	−0.02	0.03	0.01	0.00
MVCCAR(t)tr.temp.mv	14234	12915	1319	fatal and serious	−0.09	−0.08	0.00	−0.09	−0.02
				slight	0.00	−0.03	0.03	0.00	−0.04

are gradually improved from model PL to model PLtr-re by including the trend variable and the metropolitan county effect variables. The DIC is greatly improved by adding random effects in the models. Model CCAR(t)tr.temp includes both the spatial effects and the temporal effects and assumes that the spatial random effects are independent for accidents of different severity and are different over time. This has two implications. Firstly, it indicates that the extent of spatial correlation in neighbouring local authorities for two types of accident (fatal and serious accidents, and slight accidents) is different. Secondly, it suggests that spatial random effects are not constant over time. Model MVC-CAR(t)tr.temp.mv takes account of the correlation between two types of accidents by using a multivariate normal prior to model the correlation between the two unstructured components and using a multivariate CAR prior to model the correlation between the two spatial components. Its DIC is the lowest and reduces the DIC of model CCAR(t)tr.temp that does not take account of the correlation between different types of accidents by 40. However, compared with the level of DIC that is over 10,000, this reduction is not much.

Without the inclusion of extra random effects in the models (see models PL and PLtr-re), Moran's I in the residuals for both types of accidents is significant from 2001 to 2003. However, Moran's I is not significant in most cases for the years 2004 and 2005. This indicates that the extent of the spatial correlation in the residuals could vary with time. This suggests that different variance parameters might be used to formulate the CAR prior in different years or in different periods. When the same variance parameter was used for the CAR prior in all the study period, for some parameters, convergence of the MCMC iterations was found to be very poor. After applying different variance parameters, all the parameters appear to converge satisfactorily as indicated by $\hat{R}$ (see Appendix C). Values of Moran's I on residuals from the CAR models are all nonsignificant. Moreover, the spatial correlation in residuals from model PLNtr in most years, that includes log-normal random effects, is also found to be nonsignificant. But the DIC of this model is approximately 5000 larger than the DICs of the CAR models.

Table 6.14 shows the estimated variance τ_θ for the spatial component in a CAR model (model CCAR(t)tr.temp). The variance for fatal and serious accidents is larger than that

for slight accidents. For both types of accidents, the variance is likely to drop over time. The posterior mean of variance parameter τ_ε for the unstructured effects is 0.02, which is small compared with the variance of spatial random effects for fatal and serious accidents. As used previously, ratio of τ_θ and τ_ε reflects the relative strength of the spatial random effects against the unstructured heterogeneity. The average weight, equivalent to the average number of neighbours, for each local authority, is 4.9 (length of neighbours list (728)/number of local authorities (149)). The ratios for fatal and serious accidents vary from 5 to 3 over time and are equal or less than 2 for slight accidents. This suggests that for fatal and serious accidents the spatial heterogeneity and the unstructured heterogeneity have similar strength while for slight accidents the unstructured heterogeneity dominates the spatial heterogeneity.

Table 6.14: Summary of the variance parameter τ_θ for the spatial component in a CAR model

Model	Severity	2001	2002	2003	2004	2005
CCAR(t)tr.temp	Fatal and serious	0.10	0.09	0.07	0.06	0.06
	Slight	0.04	0.04	0.03	0.02	0.02

Model MVCCAR(t)tr.temp.mv takes account of the within-area (conditional) correlation between the two unstructured components of variation (one for each response variable) and between the spatial components of variation. The conditional correlation implied by the multivariate normal prior is 0.42. Different variance parameter is used for the spatial component in different years. The conditional correlation between the spatial components for the five years are estimated to be 0.78, 0.78,0.63,0.69 and 0.54. This indicates that the correlation between the spatial components is fairly large. For the formula used to calculate the correlation, see Section 4.5.

6.3.3 Maps of spatial effects

In the previous section, when accident data in the earlier years were studied, clusters of residuals with similar values were found in a residual map if a non-CAR model was used. After a CAR model was used to take account of the spatial random effects, no apparent

clusters of residuals were exhibited. The formulation of a CAR model implies that the spatial random effects should be spatially correlated. In other words, they should have similar values in neighbouring areas. Maps of spatial random effects are expected to show apparent clusters. Such maps can show the influence of the spatial effects on the expected number of accidents in each area according to the sign and the level of them.

During model estimation, the spatial random effect θ_{it} in area i in year t in a CAR model can be estimated and saved in each iteration. Therefore, the posterior distribution of θ_{it} can be obtained. Figure 6.23 illustrates the 95% credible intervals of the spatial random effects estimated from model CCAR(t)tr.temp in each local authority in 2001. It shows that almost all the credible intervals of the spatial random effects in London boroughs contain only positive values for both fatal and serious accidents, and slight accidents. Districts in South Yorkshire, West Yorkshire, Merseyside, Great Manchester and Tyne and Wear are more likely to have negative spatial effects for both types of accidents. There are also other local authorities in which the 95% credible intervals of the spatial effects contain only positive or negative values. However, they are difficult to be identified in such a figure. In this case, a map of spatial effects is more useful.

Figures 6.24 and 6.25 show maps of the spatial effects that were estimated from model CCAR(t)tr.temp. Local authorities in which the 95% intervals of the spatial effects contain only positive values or negative values are plotted using two different colours. Other local authorities are in white. The figures do not show the level of the spatial effects but only show the sign of them. They give a quick look at the areas that have positive or negative spatial effects. Since the spatial effects are assumed to vary over time in model CCAR(t)tr.temp, different values of these effects were obtained in different years. Considering only their signs,they are found to have similar distribution in the maps in the first three years (2001-2003) as well as in the last two years (2004-2005). Therefore, maps of the spatial effects for each type of accident in 2001 and in 2005 are given here.

Signs of the spatial effects suggest the relative influence of the unobserved or unmeasured contributory factors on the accident frequencies. Such factors are assumed to be spatially correlated and are taken account of by the CAR prior. One implication of this

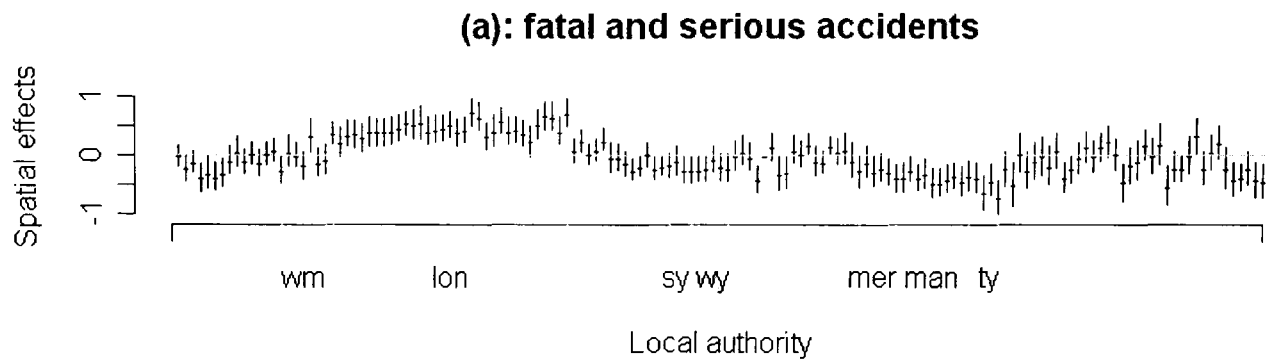

Figure 6.23: 95% credible intervals of the spatial effects in model CCAR(t)tr.temp in 2001: 'wm' for West Midlands; 'lon' for London boroughs; 'sy' for South Yorkshire; 'wy' for West Yorkshire; 'mer' for Merseyside; 'man' for Great Manchester; 'ty' for Tyne and Wear.

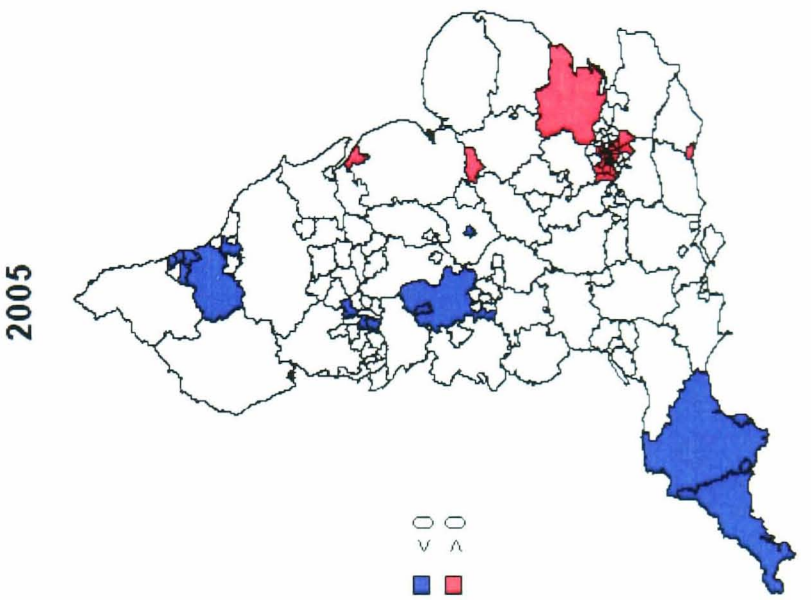

Figure 6.24: England maps of the spatial effects for model CCAR(t)tr.temp: fatal and serious accidents.

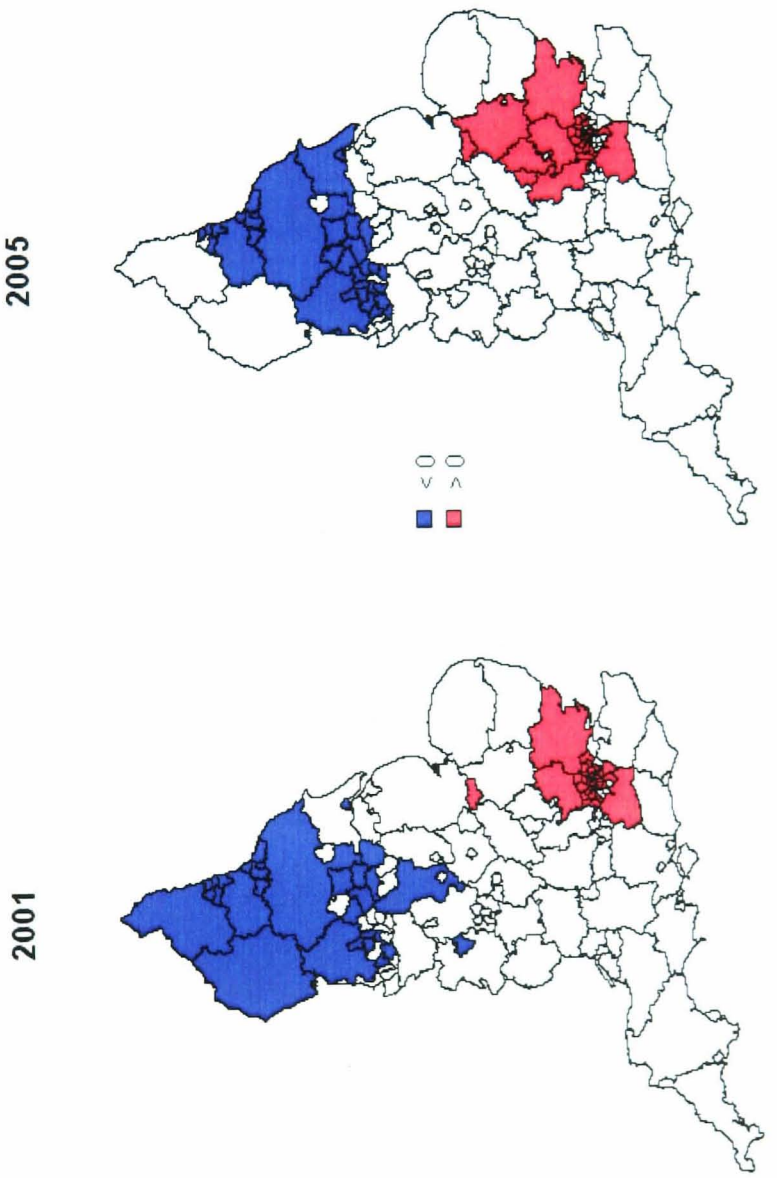

Figure 6.25: England maps of the spatial effects for model CCAR(0)tr.temp: slight accidents.

is that the expected number of accidents in a local authority depends on not only the explanatory variables but also the level of the spatial effect. A positive spatial effect in an area is associated with a larger expected number of accidents in the area compared with an area which has similar values for the explanatory variables and a zero or negative spatial effect. Similarly, a negative spatial effect in an area is associated with a smaller expected number of accidents in the area compared with an area which has similar values for the explanatory variables and a zero or positive spatial effect.

The estimated medians of the level of spatial effects are not as large as those found in spatial epidemiology. Richardson et al. (2006) studied the level of spatial effects modelled by the shared component CAR models. The estimated medians of the spatial effects in their studies are over 1 in many locations. This reflects the strong influence of the spatial effects on disease rates. In disease mapping, except for using population to adjust the relative disease risk in an area, models usually do not include explanatory variables. Therefore, the variation in the response variable is mainly explained by random effects like spatial effects. This could be the reason for obtaining larger estimates of the spatially structured effects in Richardson et al. (2006) than those found in this research.

6.3.4 Temporal correlation

For models without temporal correlation, high correlation is identified in the residuals for different time periods. The first order temporal correlation in the residuals is around 0.85 for both types of accidents as shown in tables 6.15 and 6.16.

After using a first order autoregressive prior to take account of the temporal effects, the temporal correlation in the residuals drops for both types of accidents as shown in Tables 6.17 and 6.18. The above analysis of the temporal correlation in the residuals indicates that the first order autoregressive prior has successfully explained the temporal correlation in the residuals from models that do not include the temporal random effects.

Table 6.15: Temporal correlation coefficients for residuals from model CCAR(t)tr for fatal and serious accidents

Year	2001	2002	2003	2004	2005
2001	1	0.88	0.85	0.81	0.77
2002	0.88	1	0.87	0.80	0.73
2003	0.85	0.87	1	0.85	0.76
2004	0.81	0.80	0.85	1	0.86
2005	0.77	0.73	0.76	0.86	1

Table 6.16: Temporal correlation coefficients for residuals from model CCAR(t)tr for slight accidents

Year	2001	2002	2003	2004	2005
2001	1	0.89	0.84	0.78	0.75
2002	0.89	1	0.88	0.83	0.75
2003	0.84	0.88	1	0.86	0.80
2004	0.78	0.83	0.86	1	0.86
2005	0.75	0.75	0.80	0.86	1

Table 6.17: Temporal correlation coefficients for residuals from model CCAR(t)tr.temp for fatal and serious accidents

Year	2001	2002	2003	2004	2005
2001	1	0.26	0.07	0.08	0.01
2002	0.26	1	−0.07	−0.14	−0.04
2003	0.07	−0.07	1	−0.43	−0.15
2004	0.08	−0.14	−0.43	1	−0.60
2005	0.01	−0.04	−0.15	−0.60	1

Table 6.18: Temporal correlation coefficients for residuals from model CCAR(t)tr.temp for slight accidents

Year	2001	2002	2003	2004	2005
2001	1	0.35	0.07	−0.02	0.01
2002	0.35	1	−0.23	−0.02	−0.05
2003	0.07	−0.23	1	−0.48	−0.06
2004	−0.02	−0.02	−0.48	1	−0.64
2005	0.01	−0.05	−0.06	−0.64	1

6.3.5 Estimated coefficients

Explanatory variables included in the models are area, population, length of A-roads, length of B-roads, length of minor roads, traffic by cars, traffic by other vehicles and number of junctions (represented by nodes). Accidents occurs because there is traffic. Traffic is generated by people and is carried by roads of different classes. Therefore, accident frequency is expected to be positively associated with level of traffic, population

and road length. Moreover, suppose that the total length of roads in a local authority is fixed. The road network will become more complicated if there are more junctions. So a positive correlation is also expected between accident frequency and number of junctions.

Figures 6.26 to 6.29 show the estimated medians and 95% credible intervals for the coefficients of all the explanatory variables in models PLNtr (a Poisson linear model with log-normal random effects and with a linear trend) and CCAR(t)tr.temp (a CAR model with temporal effects formulated by a first order autoregressive prior) for the two types of accidents. There is smaller variation in the estimates of the coefficients from model PLNtr than that from the CAR model. For fatal and serious accidents, the main difference in the coefficient estimates between two models is found to be with two variables, length of minor roads and B-roads. The credible interval of the coefficient for variable length of minor roads in model PLNtr contains only negative values but contains only positive values in model CCAR(t)tr.temp. Therefore, different relationship between length of minor roads and the expected number of fatal and serious accidents is obtained from the two models. For model PLNtr, the credible interval of the coefficient for variable length of B-roads covers both positive and negative values and its median is close to zero. This indicates that this variable does not contribute much to explain the variation in fatal and serious accidents at the local authority level. However, in model CCAR(t)tr.temp, the credible interval of the coefficient for variable length of B-roads contains only positive values. Therefore, it indicates that this variable is useful to account for the variation in fatal and serious accidents. Comparing with the expected relationship between the response variables and variables for road length discussed earlier, results from the CAR model are more reasonable. The credible intervals for variables, namely population, length of A-road, traffic by other vehicles and number of node, contain only positive values. The result suggests that a larger number in one of these variables holding others constant will be associated with a larger expected number of fatal and serious accidents at the local authority level. The credible intervals for variables, namely area and traffic by cars, contain only negative values for both of the models. This suggests that there is a negative relationship between the expected number of fatal and serious accidents with the two variables.

6.3 Models for accidents at the local authority level in England from 2001 to 2005

For slight accidents, the credible intervals of the coefficients for variables length of B-roads and length of minor roads contain some negative values in model PLNtr but contain only positive values in model CCAR(t)tr.temp. These again suggest that estimates from the CAR model are more reasonable. As found in models of fatal and serious accidents, there is a negative relationship between the expected number of slight accidents with area and traffic by cars.

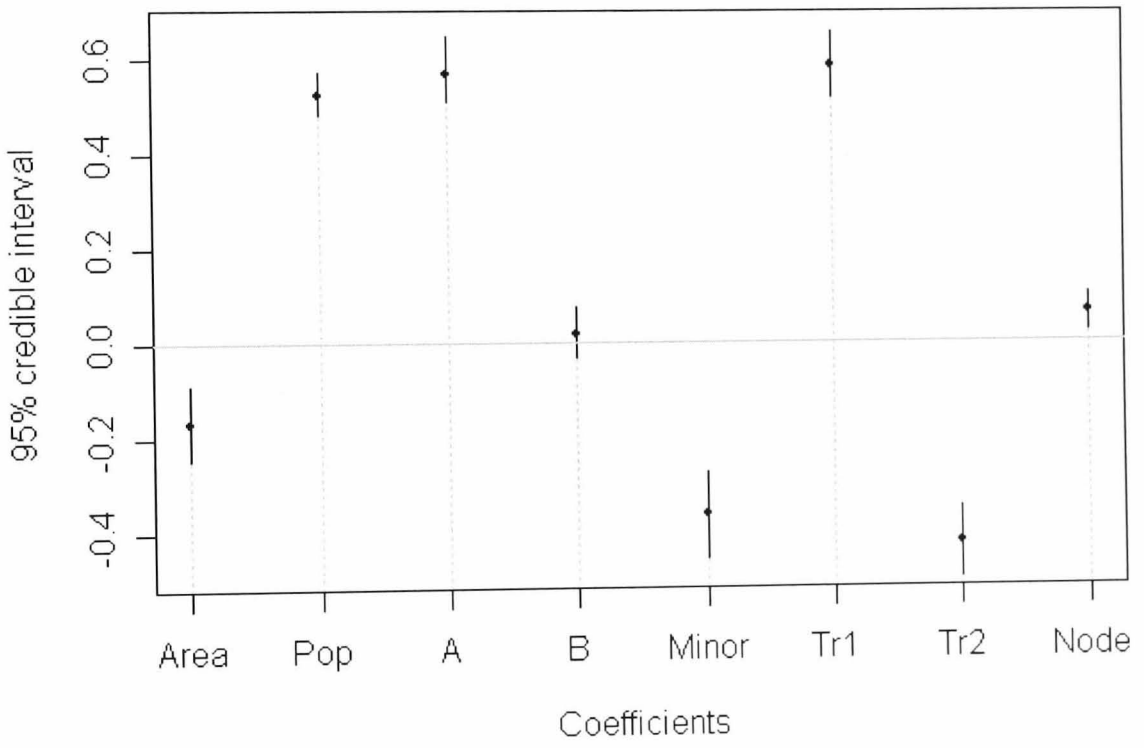

Figure 6.26: Credible intervals of the coefficients in model PLNtr for fatal and serious accidents: explanatory variables from the left to the right are area, population, length of A-roads, length of B-roads, length of minor roads, traffic by other vehicles, traffic by cars and number of nodes.

A positive correlation between the expected number of accidents with road length, population and number of junctions is consistent with the expectation. However, for the two traffic variables, a negative coefficient is obtained for car traffic and a positive one is obtained for traffic of other vehicles. Suppose that the coefficient for traffic of other vehicles is $a\,(a > 0)$ and the coefficient for car traffic is $-b\,(b > 0)$. If only these two variables, denoted by x_1 and x_2, are included in the model and the constant term is

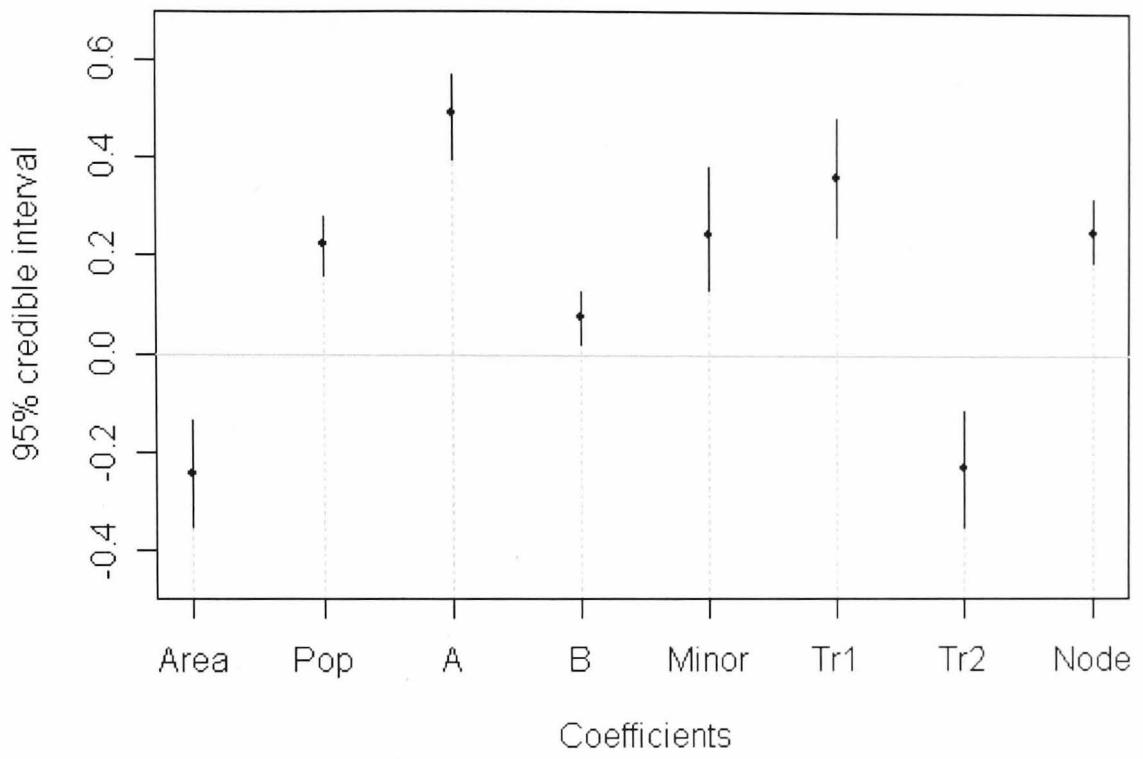

Figure 6.27: Credible intervals of the coefficients in model CCAR(t)tr.temp for fatal and serious accidents: same explanatory variables as in Figure 6.26.

ignored, we will have

$$\log \lambda = a \log x_1 - b \log x_2.$$

After taking anti-log, the following equation can be obtained:

$$\lambda = (x_1)^a (x_2)^{-b}. \tag{6.2}$$

Suppose the total traffic is $T = x_1 + x_2$ and the percentage of traffic by other vehicles is $p\,(0 < p < 1)$. Therefore, $x_1 = Tp$ and $x_2 = T(1-p)$. Equation (6.2) can be rewritten as

$$\begin{aligned} \lambda &= (Tp)^a [T(1-p)]^{-b} \\ &= T^{(a-b)} p^a (1-p)^{-b}. \end{aligned} \tag{6.3}$$

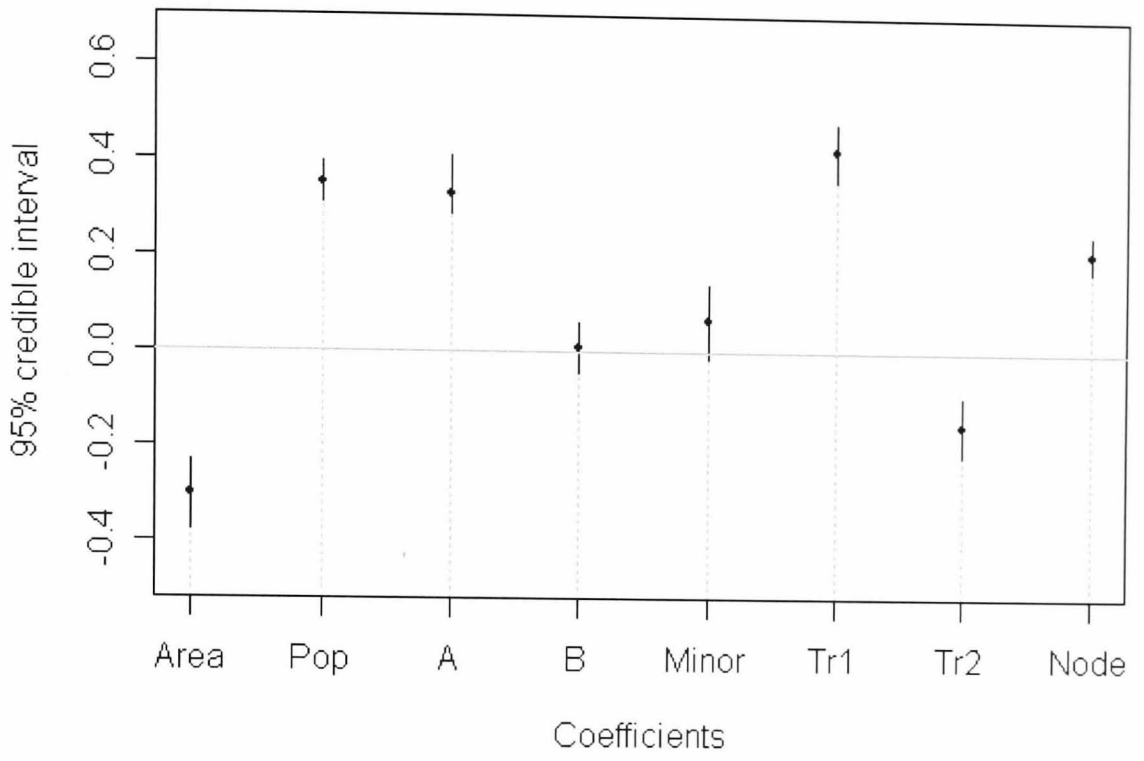

Figure 6.28: Credible intervals of the coefficients in model PLNtr for slight accidents: same explanatory variables as in Figure 6.26.

The credible intervals in the figures show that the absolute value of the coefficient for traffic of other vehicles is larger than that for car traffic. Therefore, $a - b > 0$. Equation (6.3) suggests that there is positive correlation between the expected number of accidents λ and the total traffic T. However, there is also a relationship between λ and $f(p) = p^a(1-p)^{-b}$, a function of the percentage of traffic of other vehicles. The derivative of $f(p)$ is $ap^{(a-1)}(1-p)^{-b} + p^a b(1-p)^{(-b-1)}$. It is always larger than 1 because $a > 0$ and $b > 0$. Therefore, $f(p)$ is monotonically increasing. This suggests that when two local authorities have same level of total traffic, the local authority that has a higher proportion of traffic by other vehicles is expected to have a larger expected number of accidents than that in the other local authority.

Moreover, a downward trend was identfied for both types of accidents in models with a linear trend. For instance, in model CCAR(t)tr.temp, the posterior median, denoted by δ, for the coefficient of the trend variable for fatal and serious accidents is -0.09 with

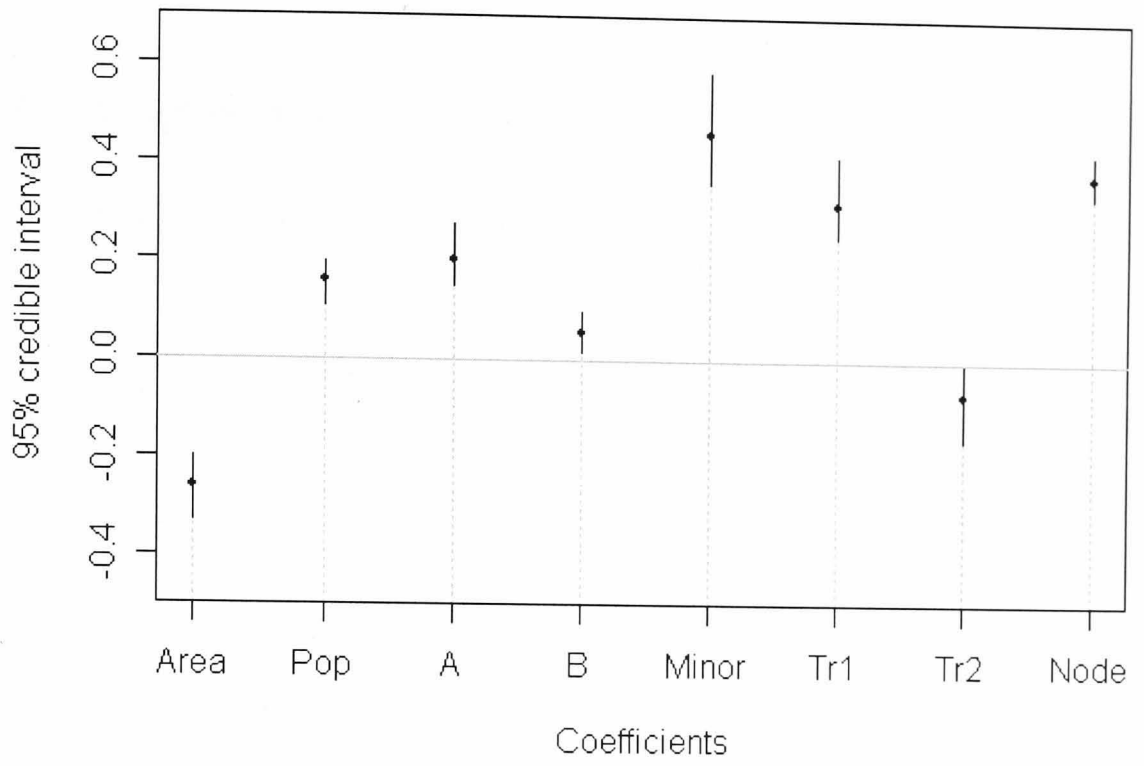

Figure 6.29: Credible intervals of the coefficients in model CCAR(t)tr.temp for slight accidents: same explanatory variables as in Figure 6.26.

the 95% credible interval to be $[-0.11, -0.06]$. For slight accidents, it is -0.05 with the 95% credible interval to be $[-0.07, -0.03]$. This may indicate that the expected number of accidents at the local authority level decreases by some rate over time. According to the form of accident models used in this thesis, the posterior median of the decreasing rate should be $\exp(\delta)$ and therefore is approxmiate to 10% for fatal and serious accidents and 5% for slight accidents. However, what this trend represents or explains is difficult to identify.

A dummy variable was also included in some models to represent the effect of unitary authorities. For fatal and serious accidents, in model CCAR(t)tr.temp, the posterior median of the coefficient for this variable is 0.12 with the 95% credible interval to be $[0.06, 0.18]$. This indicates that there are more fatal and serious accidents in an unitary authority than in other types of local authorities if they have similar values of explanatory variables. But there is no evidence that there are more slight accidents in unitary authori-

ties than in other local authorities. The 95% credible interval of the coefficent for unitary authority contains both positive and negative values.

6.4 Models for accidents at the ward level in the West Midlands in 2001

In the previous two sections, models were fitted at the local authority level. Such models can also be used to study the relationship between the accident frequency and other variables as well as the effect of the inclusion of spatial random effects at a more local level, such as ward. In this section, similar forms of models in the previous sections are fitted using the West Midlands data in 2001. Details of the response variables and explanatory variables were introduced in Section 5.3.3.

6.4.1 Relationships of the variables

Nine explanatory variables were included in the accidents models for the West Midlands. They are population, area, length of major roads, length of minor roads, number of junctions, population travelling to work by bus, population travelling to work by car as a driver, population travelling to work by car as a passenger and population travelling to work on foot. The response variables are the number of fatal and serious accidents in a ward and the number of slight accidents in a ward. Figure 6.30 shows the relationships of these variables. All the variables are in logarithmic form. The figure shows that both of the response variables are positively correlated with each explanatory variable.

6.4.2 Description of the models

Log-normal random effects, fixed metropolitan distric effects and spatially structured random effects were added in the Poisson log-linear models (PL) in turn. These models are denoted by PLN, PLNre, CCAR, MVCCAR and MVCCAR.mv. The last two models are multivariate CAR models. Three structures of the neighbours list were used. The first

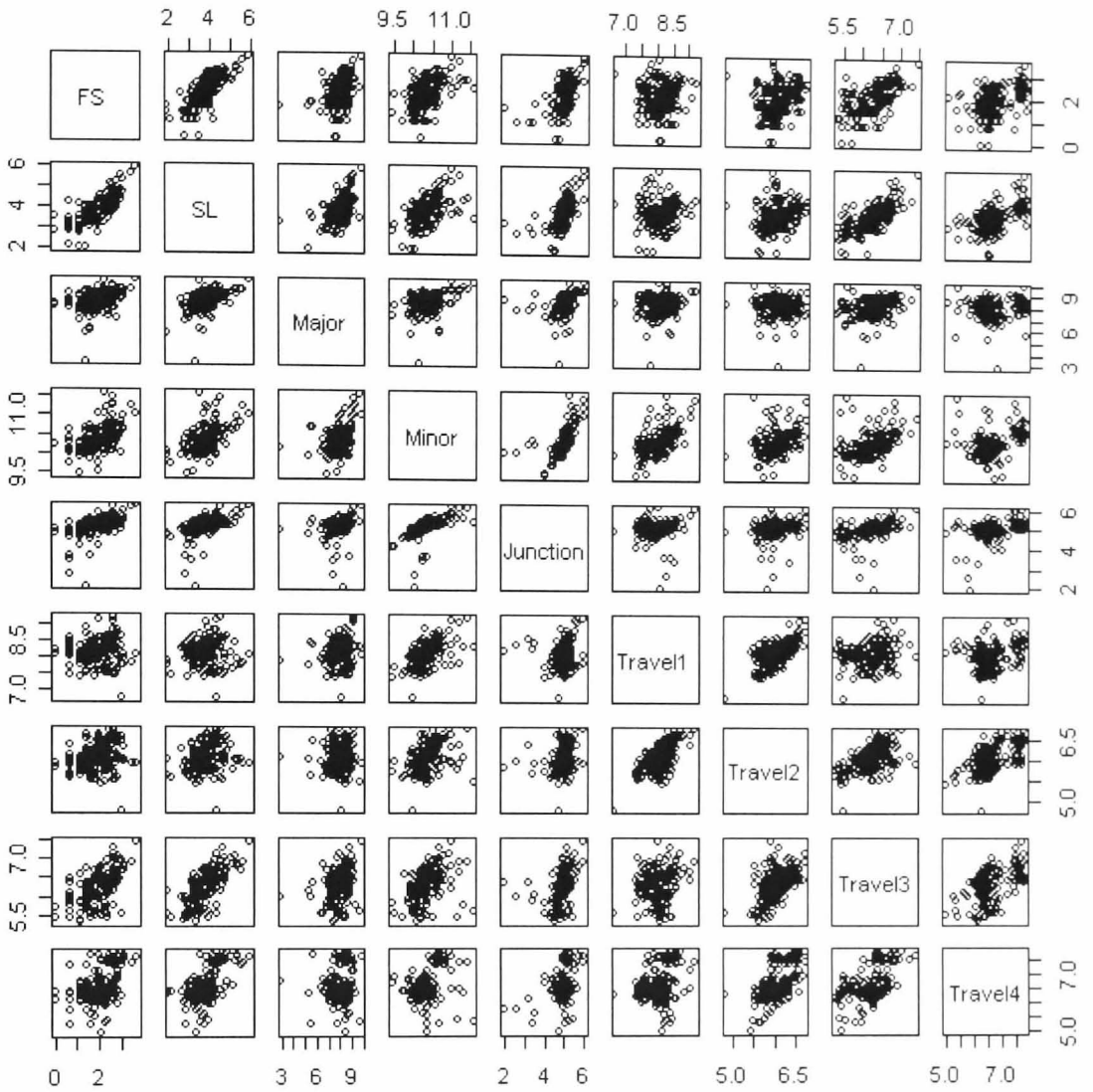

Figure 6.30: Relationships of selected variables in logarithmic form: 'FS' for the fatal and serious accidents; 'SL' for the slight accidents; 'Major' for the length of major roads; 'Minor' for the length of minor roads; 'Junction' for the number of junctions; 'Travel1' to 'Travel4' for population travelling to work by car as driver, by car as passenger, on foot and by bus respectively.

one includes only the first-order neighbours, the second one includes both the first-order neighbours and the second-order neighbours and the last one considers all the other wards in the West Midlands as neighbours. For the first order neighbours, 1-0 weights were used. For higher order neighbours, the spatial weights determined by Euclidean distance were used.

Table 6.19: Summary of the model fits for accidents in the West Midlands

Model	DIC	Expected deviance	Effective number of parameters	Severity	*Moran's I*
PLN	2034	1855	179	fatal and serious	0.11(*)
				slight	0.08
PLNre	1992	1851	141	fatal and serious	0.07
				slight	0.02
CCAR	1994	1841	153	fatal and serious	0.01
				slight	0.03
MVCCAR	1993	1841	152	fatal and serious	−0.01
				slight	−0.09
MVCCAR.mv	1998	1828	170	fatal and serious	0.00
				slight	−0.06

*: significant at the 5% level

6.4.3 Models comparison and interpretation

Table 6.19 summarises the fit of selected models. The spatial correlation in the response variables are 0.48 for fatal and serious accidents and 0.62 for slight accidents. The extent of spatial correlation in the residuals from the Poisson log-linear models with log-normal random effects, denoted by PLN in the table, is not very high for either type of accident. This indicates that the high spatial correlation in the response variables are mostly captured by the explanatory variables. However, Moran's I is still significant for fatal and serious accidents.

The inclusion of metropolitan county effects (corresponding to model PLNre) causes the value of Moran's I to drop for both types of accident and improves the DICs. By using a CAR prior to take account of the spatially structured random effects, model CCAR has a similar DIC as model PLNre. Moran's I is not significant for either type of accident and is very small. These results suggest that models with metropolitan district effects (PLNre) and models with spatial random effects (CCAR) perform similarly for this dataset. CAR models including higher order neighbours were also considered. However, they do not perform better than the one just including first order neighbours.

The variance parameter of the unstructured random effect is 0.02. For the spatially structured random effect in model CCAR, the overall variance parameter is 0.16 for fatal and serious accidents and 0.13 for slight accidents. Divided by the average number of neighbours for each ward, which is about 5.4, the variance of the spatially structured effect for each ward is about 0.03 for fatal and serious accidents and 0.02 for slight acci-

dents. Therefore, the ratio between the variances of two types of random effects is about 1.5 for fatal and serious accidents and 1 for slight accidents. This suggests that the unstructured random effects and the spatially structured random effects included in the ward models have the similar importance although for fatal and serious accidents the spatially structured effects are a little stronger.

Model MVCCAR takes account of the within-area correlation in the spatially structured random effects for the two types of accident. Its DIC is very close to models CCAR. The correlation in the spatial components for accidents of different severities is estimated to be 0.67. This suggests that the correlation in the spatially random effects for the two types of accidents, namely fatal and serious accidents and slight accidents is fairly high.

Model MVCCAR.mv takes account of the within-area correlation in the two types of accident for both spatially structured random effects and unstructured random effects. Its DIC is higher than the previous three models. But this is caused by using more parameters. Its expected deviance is the lowest among all the models. The within-area correlation between the spatial effects is about 0.68. The within-area correlation between the unstructured random effects is about 0.61.

6.4.4 Estimated coefficients

Summaries of parameter estimates for selected models are listed in Appendix D. Figures 6.31 and 6.33 plot the 95% credible intervals of the coefficients for the explanatory variables in model PLN. Figures 6.32 and 6.34 plot the credible intervals for the coefficients in model MVCCAR. There is not much difference in the credible intervals for the estimated coefficients from the two models. Results from model CAR and MVCCAR.mv are also very similar.

For fatal and serious accidents, all the medians of the coefficients are positive except for the one for population travelling to work by car as driver. However, the credible interval of the coefficient for this variable covers both positive and negative values and its median is close to zero. The same result is found for the coefficient of length of minor roads except that its median is positive but also close to zero. These indicate that

population travelling to work by car as driver and length of minor roads do not contribute much to explain the variation in fatal and serious accidents at the ward level. For other explanatory variables, the credible intervals for the coefficients mostly contain positive values. These variables are population, area, length of major roads, number of nodes, population travelling to work by bus, population travelling to work by car as passenger and population travelling to work on foot. The result suggests that a larger number in one of these variables holding others constant will be associated with a larger expected number of fatal and serious accidents at the ward level.

For slight accidents, the medians of the coefficients for variables, namely length of minor roads, number of nodes, population travelling to work by car as driver and population travelling to work by bus are negative. For the first three variables, the credible intervals for the coefficients contain only negative values. This suggests that a larger number in one of these variables holding others constant will be associated with a smaller expected number of slight accidents at the ward level. For other variables including population, area, length of major roads, population travelling to work by car as passenger and population travelling to work on foot, the credible intervals for the coefficients mostly contain positive values. This suggests that a larger number in one of these variables holding others constant will be associated with a larger expected number of slight accidents at the ward level.

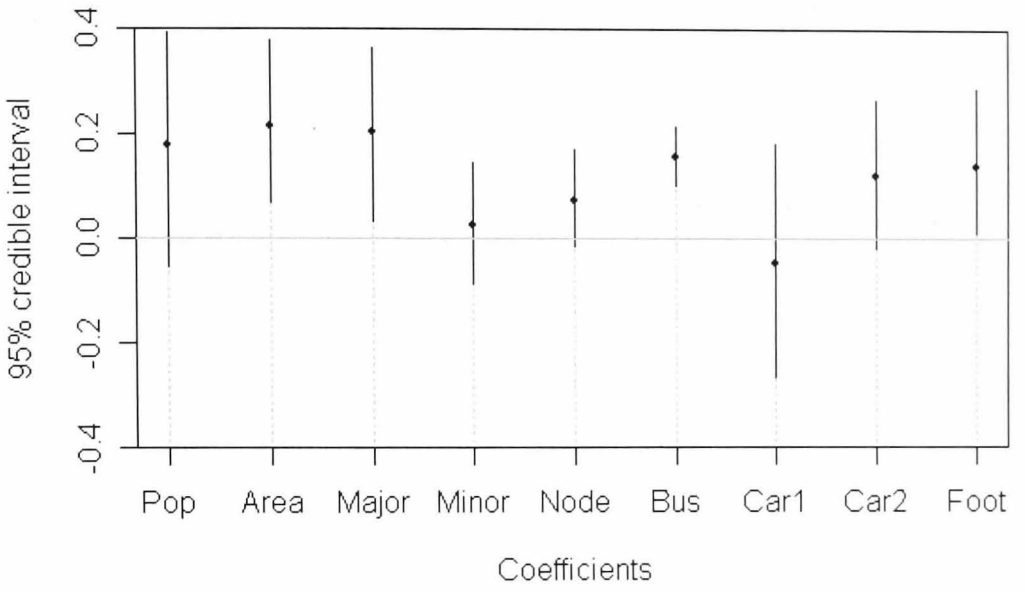

Figure 6.31: Credible intervals for the coefficients of the explanatory variables in model PLN for fatal and serious accidents: explanatory variables are in turn population, area, length of major, length of minor, number of nodes, population travelling to work by bus, by car as driver, by car as passenger and on foot respectively.

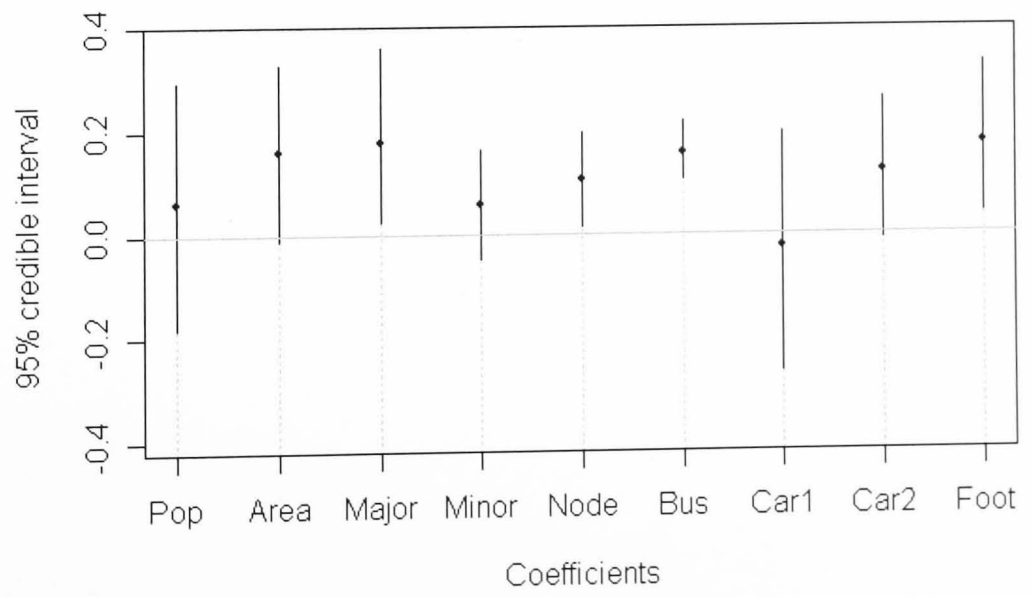

Figure 6.32: Credible intervals for the coefficients of the explanatory variables in model MVCCAR for fatal and serious accidents: same explanatory variables as in Figure 6.31.

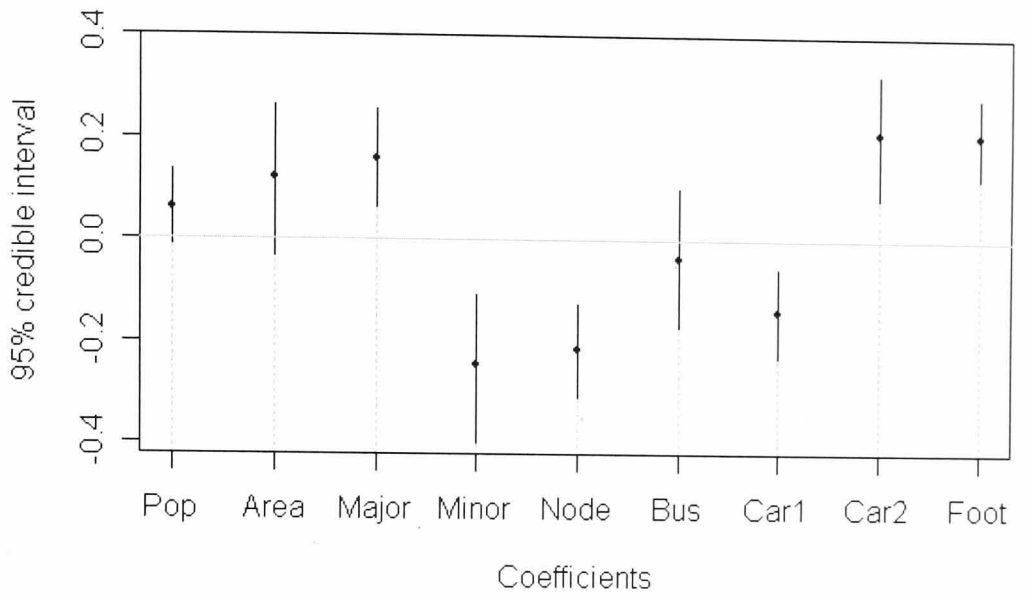

Figure 6.33: Credible intervals for the coefficients of the explanatory variables in model PLN for slight accidents: same explanatory variables as in Figure 6.31.

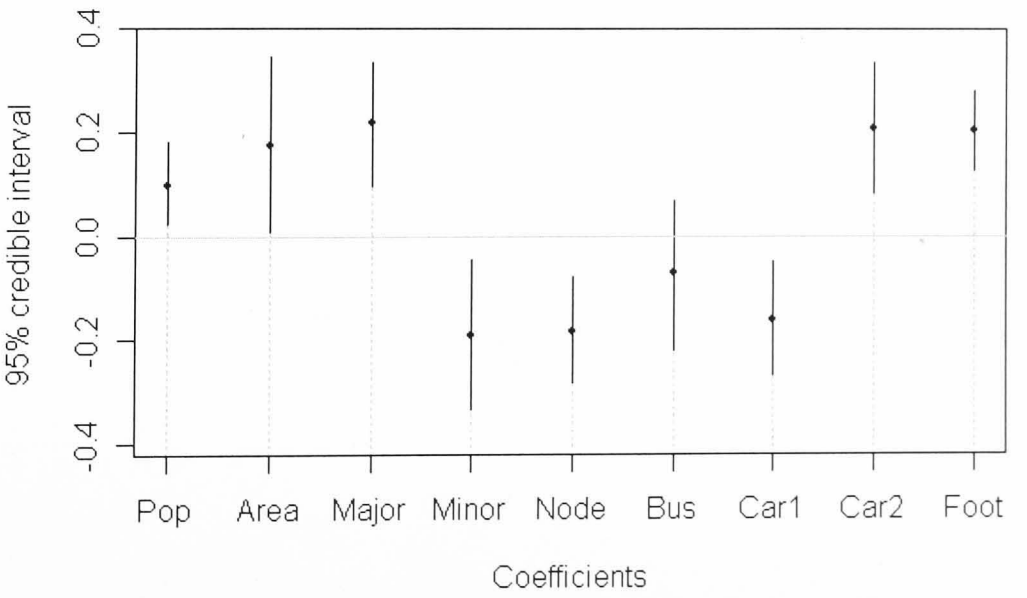

Figure 6.34: Credible intervals for the coefficients of the explanatory variables in model MVCCAR for slight accidents: same explanatory variables as in Figure 6.31.

6.4.5 More on the spatial effects

Figures 6.35 and 6.36 plot maps of the posterior medians of the spatially structured random effects in model MVCCAR. Results show that, for fatal and serious accidents, the three districts lying in the east of the West Midlands, namely Birmingham, Solihull and Coventry from left to right, have positive spatial effects in most of their wards. The spatial effects corresponding to wards in the most northern districts, namely Walsall and Wolverhampton, are all negative. These consequently result in apparent clusters of positive values and negative values. The distribution pattern of the spatial effects for slight accidents looks different from that for fatal and serious accidents. Its extent of clustering is less strong. These maps can help to identify wards with higher spatial effects that are caused by some unknown or unmeasurable factors. Levels of these spatial effects reflect the relative influence of such factors on the expected number of accidents at the ward level.

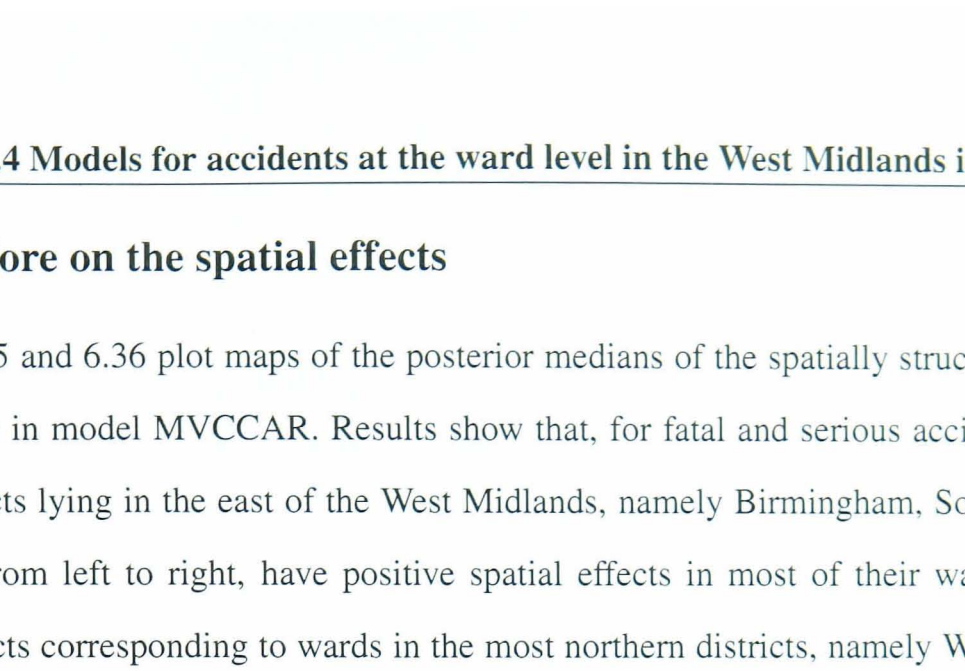

Figure 6.35: Map of the posterior medians of the spatially structured random effects in model MVCCAR: fatal and serious accidents.

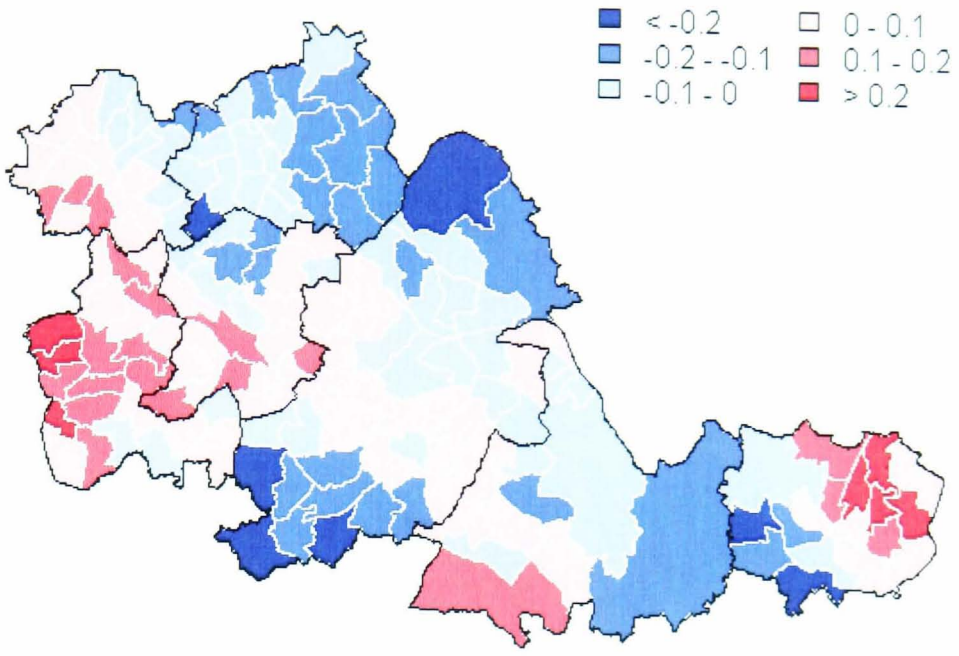

Figure 6.36: Map of the posterior medians of the spatially structured random effects in model MVCCAR: slight accidents.

6.5 More on residual spatial autocorrelation

In the previous three sections, the residuals used for examining Moran's I were Pearson residuals calculated based on the point estimates of the λs – the posterior mean of λ. As shown in Section 4.7, in a Bayesian context, a more appropriate approach is to use Bayesian residuals for model checking. Using the local authority models in the 2000s as an example, the following analysis shows how an Bayesian approach is used to examine spatial correlation in residuals. For each model, λs for 149 local authorities in 1000 simulations were saved when the models were fitted. Suppose that the estimate of Poisson mean in area i in the jth simulation is $\lambda_i^{(j)}$. Then Bayesian residuals for the jth simulation are $\frac{y-\lambda^{(j)}}{\sqrt{\lambda^{(j)}}}$. The value of Moran's I for Bayesian residuals in each simulation was examined. As shown in Section 4.7, Values of the response variable y could be predicted from $\lambda^{(j)}$ by simulating $y^{(j)}$ from the Poisson distribution Pois$(\lambda^{(j)})$. Residuals based on the predicted values of y are $\frac{y^{(j)}-\lambda^{(j)}}{\sqrt{\lambda^{(j)}}}$. Using function $I(y,\lambda)$ to represent the calculation of Moran's I for residuals, the probability that $I(y^{(j)},\lambda^{(j)}) > I(y,\lambda^{(j)})$ is a measure to check whether a model is misfitted.

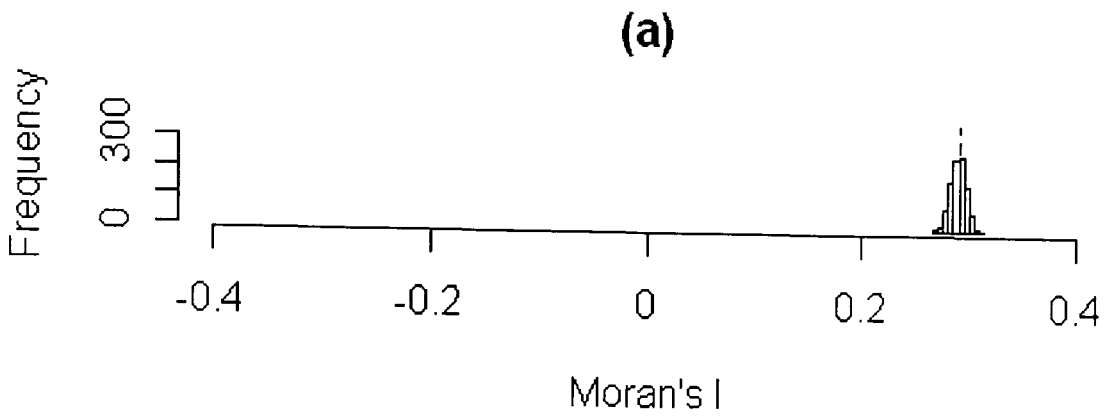

Figure 6.37: Values of Moran's I in Bayesian residuals from model PLtr: (a) based on true y; (b) based on predicted values for y.

Figures 6.37 to 6.39 show values of Moran's I on residuals for fatal and serious accidents in models PLtr, PLNtr and CCAR(t)tr.temp (see Table 6.13) from year 2001. In these figures, Histogram (a) illustrates values of Moran's I for Bayesian residuals based on the true y ($I(y, \lambda^{(j)})$), and Histogram (b) shows values of Moran's I for Bayesian residuals based on the predicted values of y ($I(y^{(j)}, \lambda^{(j)})$). The dash line in Histograms (a) corresponds to the value of Moran's I for residuals based on the posterior means of λs. Histograms (a) in Figures 6.37 and 6.38 suggest that when no spatial structured random effects are included in the model, the spatial correlation in the residuals is all positive. The spatial correlation is statistically significant in the Poisson log-linear model PLtr, but nonsignificant in model PLNtr that includes log-normal random effects. The spatial correlation is nonsignificant for residuals based on predicted y^* for both models. The two-sided p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$ for model PLtr is 0 and is 0.02 for model PLNtr.

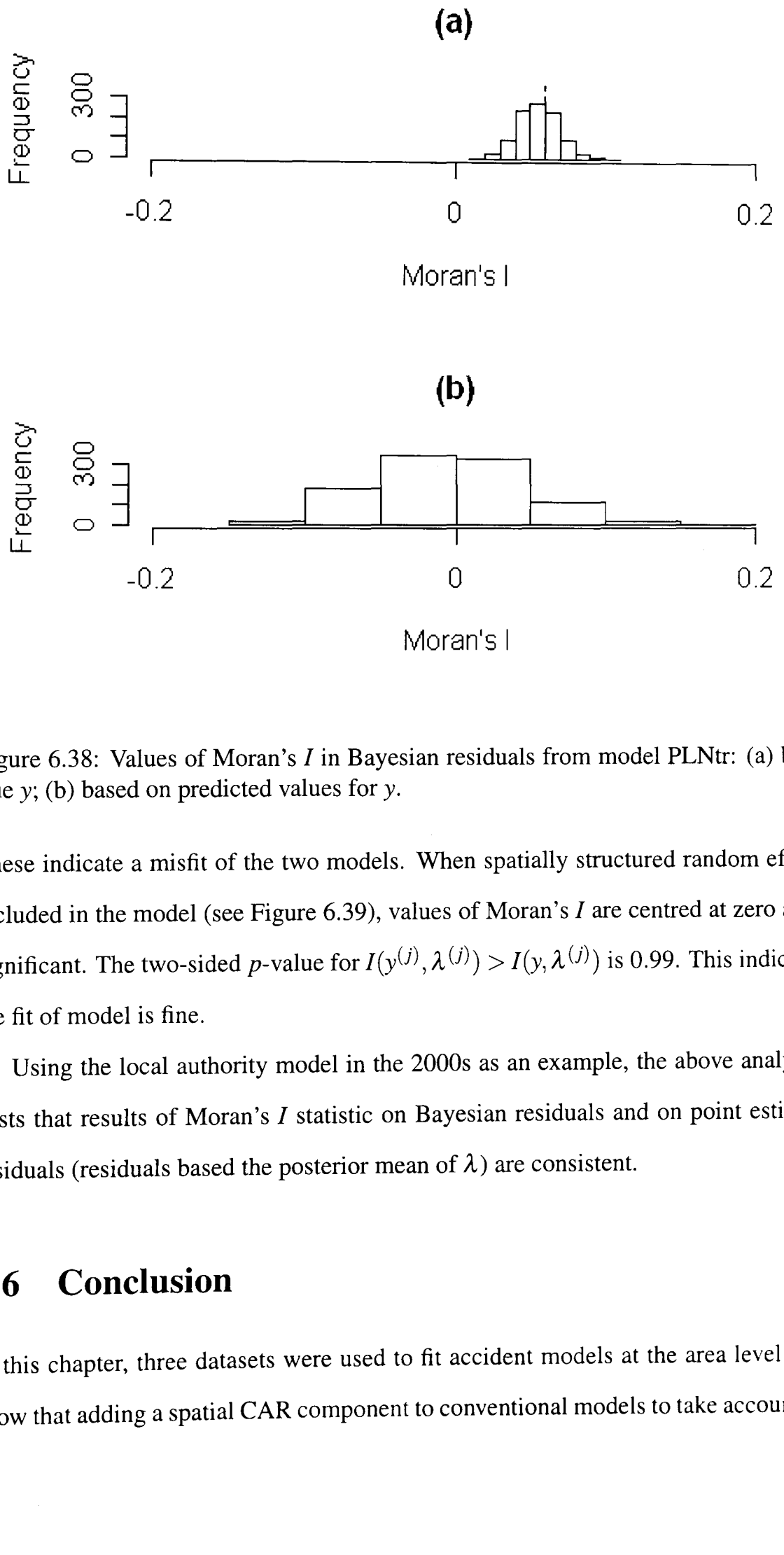

Figure 6.38: Values of Moran's I in Bayesian residuals from model PLNtr: (a) based on true y; (b) based on predicted values for y.

These indicate a misfit of the two models. When spatially structured random effects are included in the model (see Figure 6.39), values of Moran's I are centred at zero and non-significant. The two-sided p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$ is 0.99. This indicates that the fit of model is fine.

Using the local authority model in the 2000s as an example, the above analysis suggests that results of Moran's I statistic on Bayesian residuals and on point estimates of residuals (residuals based the posterior mean of λ) are consistent.

6.6 Conclusion

In this chapter, three datasets were used to fit accident models at the area level. Results show that adding a spatial CAR component to conventional models to take account of spa-

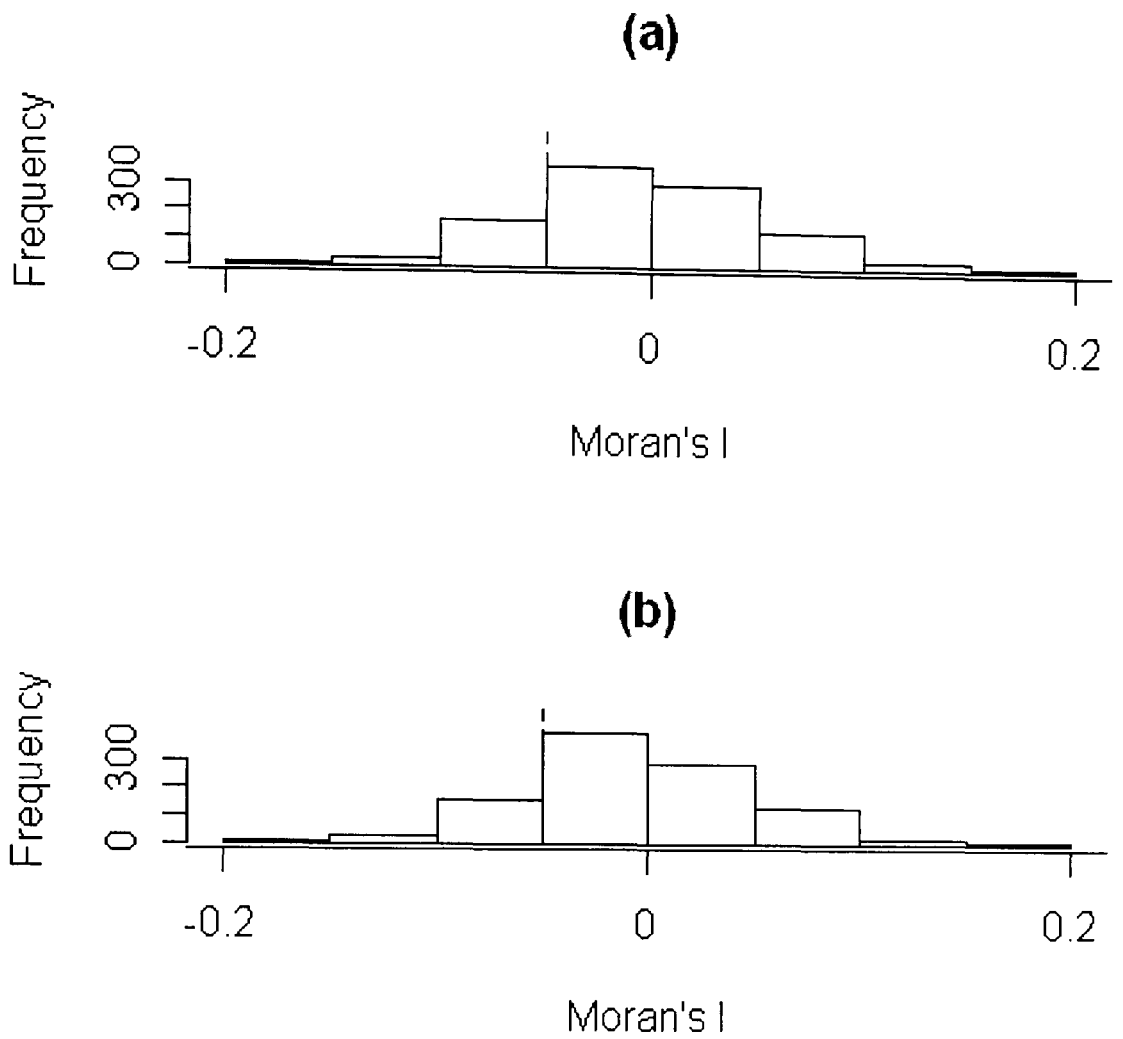

Figure 6.39: Values of Moran's I in Bayesian residuals from model CCAR(t)tr.temp: (a) based on true y; (b) based on predicted values for y.

tial random effects has two main contributions. Firstly, the model performance, measured by DIC, is improved. This indicates that the inclusion of spatial random effects better explains the variation in the response variable at the area level. Secondly, the significant spatial correlation in the posterior means of residuals is mostly removed. The spatial random effects are expected to take account of those contributory factors for accident frequencies that tend to have similar levels of values in neighbouring areas. Therefore, a map showing the spatial distribution of the spatial effects often exhibits clusters of similar values. Levels of these spatial effects reflect the relative influence of the unknown or unmeasurable contributory factors on accident frequencies.

Another finding from the CAR models at the area level is, when longitudinal data

were used, the best performing models assumed the spatial effect in an area is not constant over time. One implication from this is the unknown or unmeasurable explanatory variables captured by the spatial component may vary over time therefore their effects on the response variable change over time. However, this introduces difficulty in applying such models for predicting numbers of accidents at the area level in the future.

By taking account of the correlation in different types of accidents, multivariate models were not found to improve the models' DIC much. Their deviances are all less than those for univariate models. However, there are more effective parameters used to estimate the models. Therefore, their DICs are higher.

The shared component CAR models as introduced in Section 3.2.2 were also fitted. However, the convergence of parameters was poor and the performance of the models was not good. This indicates that such models are not appropriate for accident data used here. Normally in models for disease rates, only a limited number of explanatory variables are included. The shared spatial component for different response variables will mostly capture the effects caused by unknown or unmeasurable factors that are spatially correlated. Such effects are shared by different response variables in an appropriate form. In accident models, such effects are probably mostly captured by the available explanatory variables that are common for both types of accidents. Therefore, very little effect will be left to be taken account of by a shared CAR component. This may explain why the shared component CAR models used in disease mapping do not perform well here.

The inclusion of temporal effects using a first order autoregressive prior is found to improve the models' DIC. Such temporal effects are expected to capture the contributory factors that are not included in the models but are constant over time.

Chapter 7

Models of accidents on a road network

In the previous chapter, models with spatial and temporal effects were developed for accidents at the local authority level and at the ward level. Spatial CAR models are found to be successful to take account of the spatial random effects that are spatially autocorrelated. According to the value of DIC, CAR models perform better than non-CAR models. This chapter aims to study how spatial CAR models can be applied to road accidents on a road network and to what extent they can improve the performance of models and explain the spatial correlation in the response variable. Data analysis and models for accidents on the M1 from 1999 to 2005 are introduced first. Both spatial and temporal effects are considered. Secondly, models for junction accidents in Coventry are examined.

7.1 Models for accidents on M1

7.1.1 Some descriptive statistics

For the 59 links on the M1, both accident data and traffic data are available for 7 years. How these data were obtained and prepared have been explained in Chapter 4. Figure 7.1 shows the spread of the accident data for each year using box plots. In order to take account of the different lengths of the links, the variable used in the plot is the accident count per kilometre.

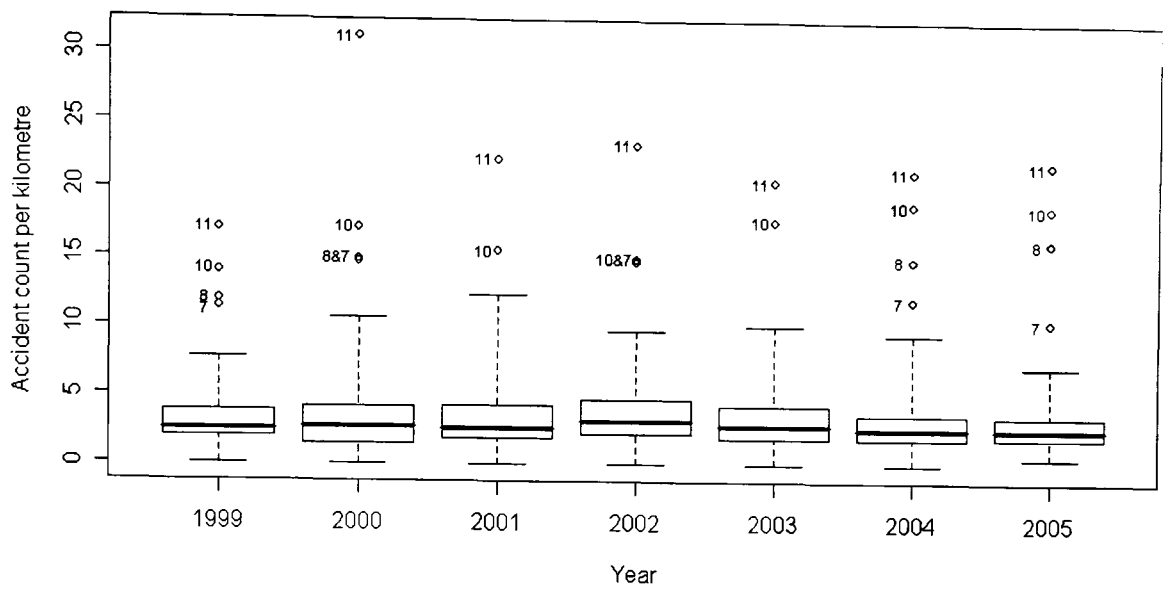

Figure 7.1: Box plots for the accident data from 1999 to 2005.

The figure indicates that for most links the accident count per kilometre is less than 10 in every year. The digits plotted beside the outliers correspond to the link IDs. Links 7, 8, 10 and 11 often appear to be outliers. They are neighbouring road links of the M1 in Hertfordshire and a very small part of Bedfordshire that intersect with the M25, M10, A4147 and A5183. There are two spurs on this road section, namely links 9 and 12 each of which has merging traffic with the A414 and A1081 respectively.

Reasons for the existence of spatial correlation in the accident data for a road network have been discussed in the previous chapters. The extent of the spatial correlation in this set of accident data for the M1 was examined. Spatial correlograms (see Section 3.3 for definition) for the accident data are shown in Figure 7.2 and 7.3. These figures show the changes of spatial correlation when the order of neighbours is increased. Neighbours up to order 10 are considered.

In order to filter out other relevant factors that may influence the value of Moran's I in accident data, link length and traffic flow are taken into account when producing the spatial correlograms over time. These correlograms show that the value of Moran's I for first order neighbours is high. Generally speaking, as the lag of neighbours increases, there is a downward trend for the Moran's I. For the accident count per kilometre, the

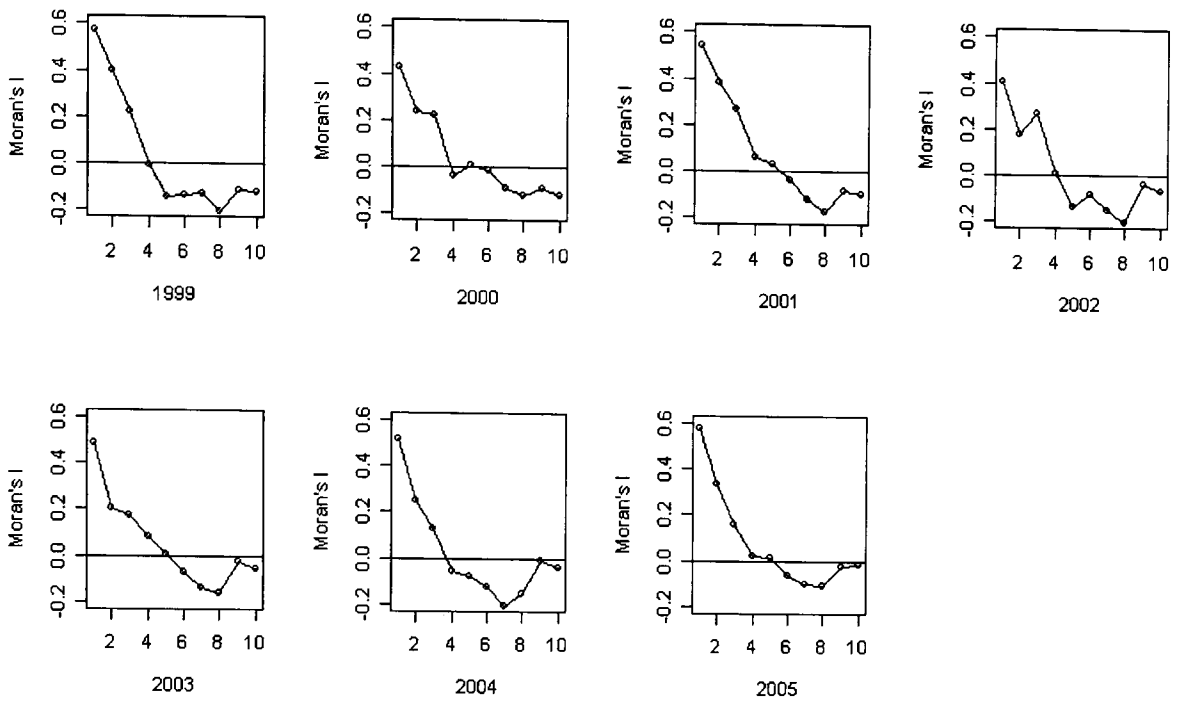

Figure 7.2: Spatial correlograms for the accident count per kilometre on the M1.

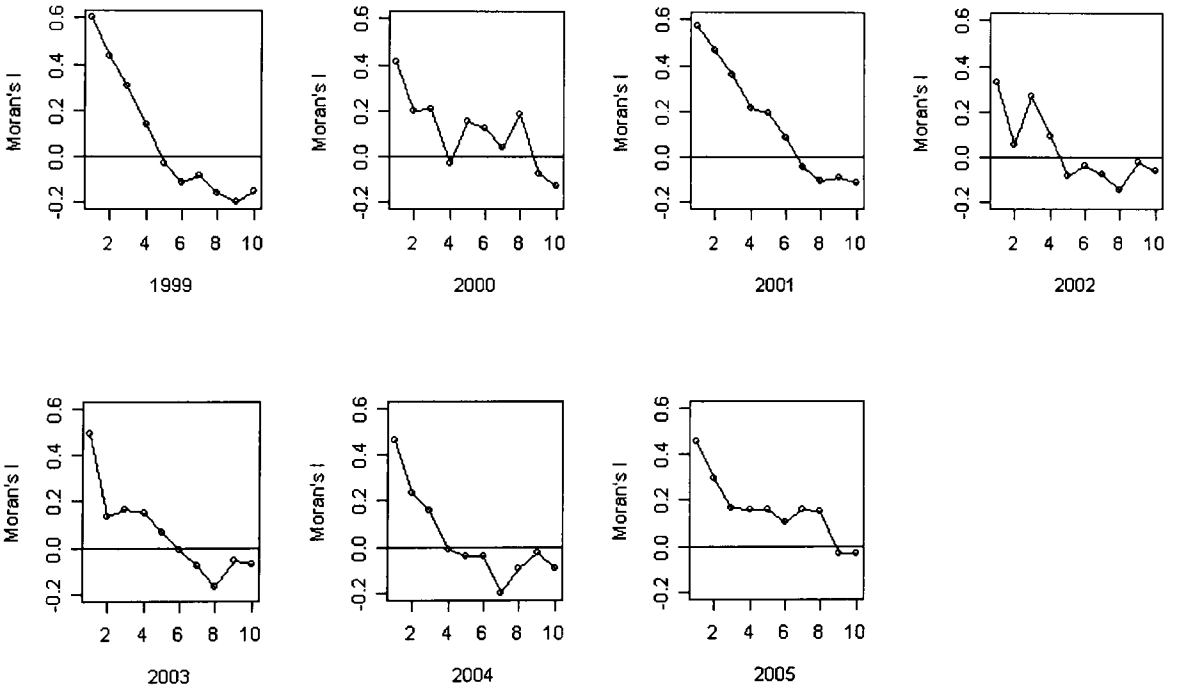

Figure 7.3: Spatial correlograms for the accident count per vehicle-kilometre on the M1.

value of Moran's I for first order neighbours is between 0.4 and 0.6. Figure 7.3 shows that the value of Moran's I for first order neighbours is still high even after taking account of the level of traffic. This indicates that the conventional models without spatial effects

may not fully explain the spatial correlation in the accident data even after including explanatory variables like traffic flow and link length. Therefore such models may not perform well and may result in spatial correlated residuals.

Similar spatial correlograms are plotted in 7.4 for the level for traffic flow. Again, the value of Moran's I for first order neighbours is fairly high. Since traffic moves on the roads, the existence of spatial correlation in the traffic level for links on a road is reasonable. For higher order neighbours, Moran's I is low in most cases.

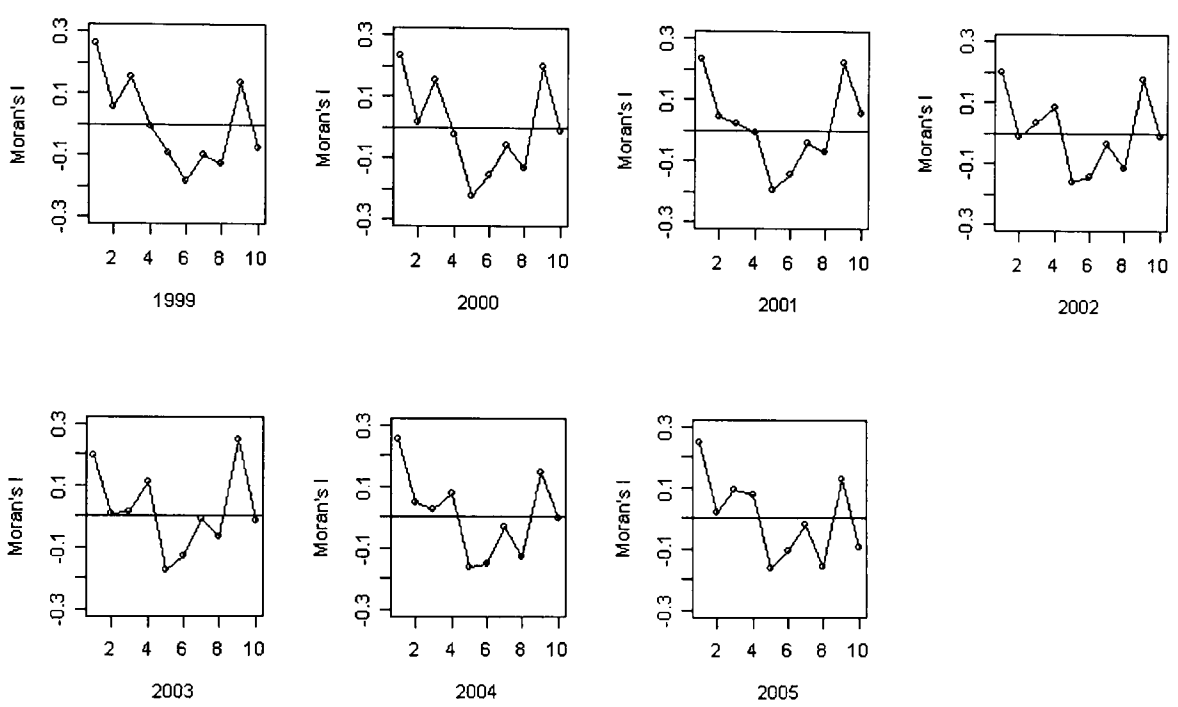

Figure 7.4: Spatial correlograms for the AADF on the M1.

Figure 7.5 shows the relationship between accident count, traffic flow and road length in 1999. All the variables are in logarithmic forms. There are some links that have no accidents. Therefore, the accident count for each link is adjusted by adding 1. The figure shows that the three variables are positively correlated.

7.1.2 Fit of the models

A number of models have been developed for the M1 data. They are models PL (Poisson log-linear model), PLN (by adding log-normal random effects to model PL), PLNtr (by adding a linear time trend to model PLN), PLNtr&re (by adding fixed regional effects to

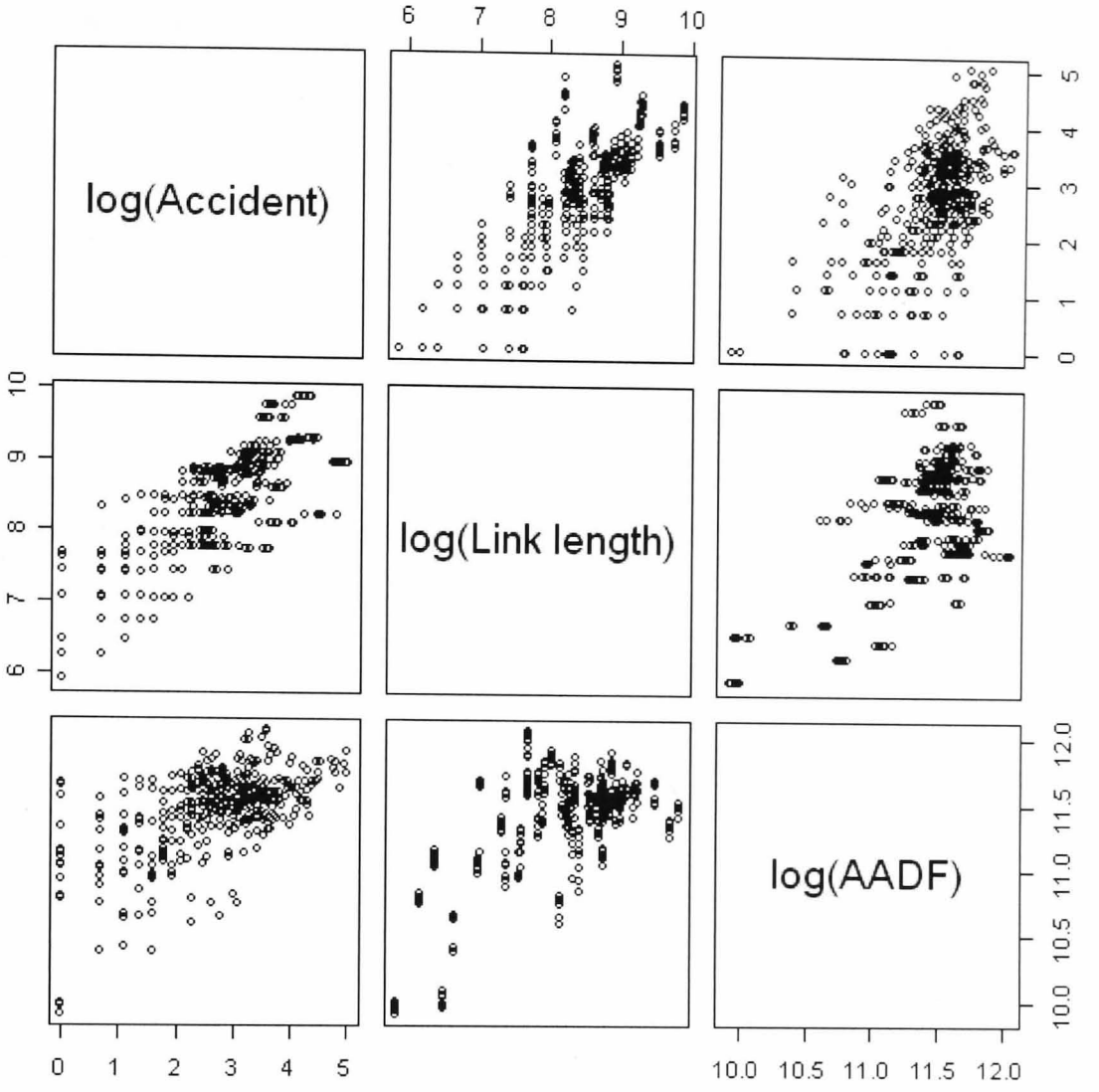

Figure 7.5: Relationship between variables

model PLNtr) and CCARtr (the convolution CAR model with a linear time trend). Two explanatory variables have been included in the models. They are the AADF and the link length. Table 7.1 shows the DIC and Moran's I statistics on the residuals for these models. According to the values of DIC in the table, the Poisson log-linear model (PL) has the

Table 7.1: Summary of the model fits for accidents on the M1 from 1999 to 2005

Model	DIC	Expected deviance	Effective number of parameters	*Moran's I*						
				1999	2000	2001	2002	2003	2004	2005
PL	4598	4595	3	0.51(∗)	0.53(∗)	0.55(∗)	0.39(∗)	0.47(∗)	0.47(∗)	0.45(∗)
PLN	2420	2132	288	0.26(∗)	0.23(∗)	0.42(∗)	0.19	0.34(∗)	0.42(∗)	0.22(∗)
PLNtr	2422	2133	289	0.24	0.25(∗)	0.42(∗)	0.19	0.35(∗)	0.40(∗)	0.19(∗)
PLNtr&re	2378	2123	255	0.13	0.07	−0.03	0.07	0.05	0.20	−0.08
CCARtr	2282	2164	117	−0.05	−0.02	−0.06	0.09	0.01	0.15	−0.02

*: significant at the 5% level

highest DIC, over 4500, therefore is the worst model. After including log-normal random effects (model PLN) in the model, the DIC has been greatly improved. Model PLNtr includes a linear time trend in model PLN. Its performance is not better than that of the model without a linear time trend. When the regions where links belong to, represented by several dummy variables, are considered in the model (PLNtr&re), the DIC is more improved. The best performing model according to the DIC is the CAR model (CARtr) that includes both the unstructured random effects and the spatially structured random effects. Values of Moran's I in the table shows that without considering any spatial effects, either fixed or random, the first three models result in spatially correlated residuals as indicated by significant Moran's I in all the years. After taking account of the spatial effects in the last two models, none of the spatial correlation in the residuals is significant. This indicates that the high spatial correlation in the residuals is successfully removed by including either the regional effects or the spatially structured random effects. Higher order neighbours were also considered to construct the neighbours list to be used in the CAR models. But according to the values of DIC, none of CAR models using higher order neighbours perform better than using the first order neighbours.

7.1.3 Estimates of the parameters

For all the models, the mean of the coefficients for the explanatory variables—AADF and link length are both positive. The mean of the coefficient for the trend variable is negative but small. This indicates that there is a downward trend in the accidents on M1. For model PLNtr&re, the mean of the coefficients for the regional effect in London, East Midlands and Yorkshire & the Humber are all negative, regarding East of England as the baseline. This indicates that there could be less accidents on M1 links that are in these regions. For model CCARtr, the variance parameter for the spatial random effect is 0.191 and is 0.013 for the unstructured random effect. The ratio of the spatial random effects against the unstructured random effects is over 14. It is much larger than the average number of neighbours for each link (length of neighbours list (120)/number of links (59) = 2). This indicates that for the M1 data the spatially structured heterogeneity dominates the

unstructured heterogeneity.

7.1.4 More on residual spatial autocorrelation

In the previous chapter, values of Moran's I in Bayesian residuals were obtained for the local authority model. The result is consistent with that found in the point estimates of residuals (residuals calculated from the posterior mean of λ). A similar approach is used here to examine the spatial correlation in Bayesian residuals and to check the fit of models. λs for 59 links in 1000 simulations were saved when the models were fitted. Suppose that the estimate of Poisson mean in area i in the jth simulation is $\lambda_i^{(j)}$. Then Bayesian residuals in the jth simulation are $\frac{y-\lambda^{(j)}}{\sqrt{\lambda^{(j)}}}$. Spatial correlation should then be examined for residuals in each simulation.

Without taking account of spatial dependency in neighbouring links, residuals from a non-CAR model were expected to be spatially autocorrelated. The extent of spatial autocorrelation in Bayesian residuals from model PLN (Poisson log-linear model with log-normal random effects) was firstly examined. The following analysis and discussion use residuals in year 1999 as an example. Figure 7.6 is a histogram of values for Moran's I in the 1000 simulations. It shows that most values of Moran's I fall in the interval between -2 and 2 and are statistically nonsignificant. However, as shown in Table 7.1, the value of Moran's I in the Pearson residuals, based on the posterior mean of λ_i, is 0.26 (illustrated by the dash line in the figure) and is significant. These suggest that contrary conclusions may be drawn when different types of residuals are used for examining spatial correlation. This result is different from what has been found for the local authority model in the previous chapter. Reasons for this are not clear at this stage and needs further investigation. Since a Bayesian approach is adopted in this research, results based on Bayesian residuals are more reliable. However, if no spatial correlation exists in Bayesian residuals from model PLN, this will suggest that the spatial correlation in the response variables has been fully explained by the model. In such a case, would a CAR model be still needed?

Model PLN includes log-normal random effects that are not spatially structured. These

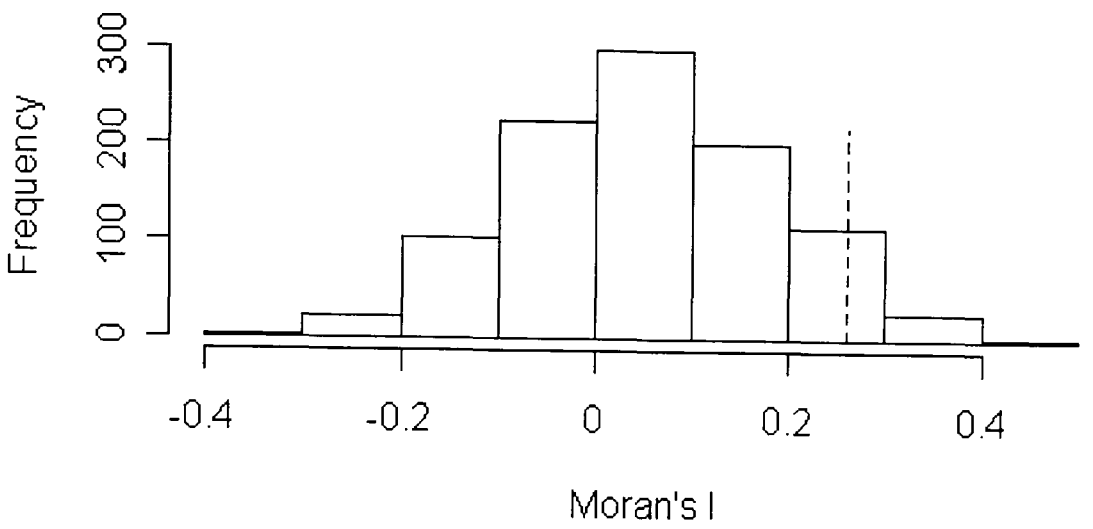

Figure 7.6: Values of Moran's I in Bayesian residuals from model PLN.

unstructured random effects capture the heterogeneity of individual site and therefore explain some extent of spatial variation. In order to examine the level of these effects, simulated values of them were saved. Figure 7.7 illustrates the 95% credible intervals of these random effects. It shows that the posterior median of the random effects for many links is close to zero. Moreover, there are a number of links whose credible intervals for the random effects contain only positive values and some of these links are neighbouring links. This figure suggests that some level of spatial variation over the links has been taken account of by the unstructured random effects. This could be a reason for obtaining non-significant Moran's I on Bayesian residuals. Therefore it suggests examining the spatial correlation in residuals from a Poisson log-linear model without extra random effects.

Values of Moran's I for Bayesian residuals from model PL in Table 7.1 are straightforward to be obtained. Moreover, values of the response variable y could be predicted from $\lambda^{(j)}$ by simulating $y^{(j)}$ from the Poisson distribution Pois$(\lambda^{(j)})$. Therefore, residuals based on the predicted values of y are $\frac{y^{(j)}-\lambda^{(j)}}{\sqrt{\lambda^{(j)}}}$. Using function $I(y,\lambda)$ to represent the calculation of Moran's I for residuals, the probability that $I(y^{(j)},\lambda^{(j)}) > I(y,\lambda^{(j)})$ for selected models could be obtained and suggests whether a model is misfitted. Figure 7.8 shows these results. Histogram (a) illustrates values of Moran's I for Bayesian

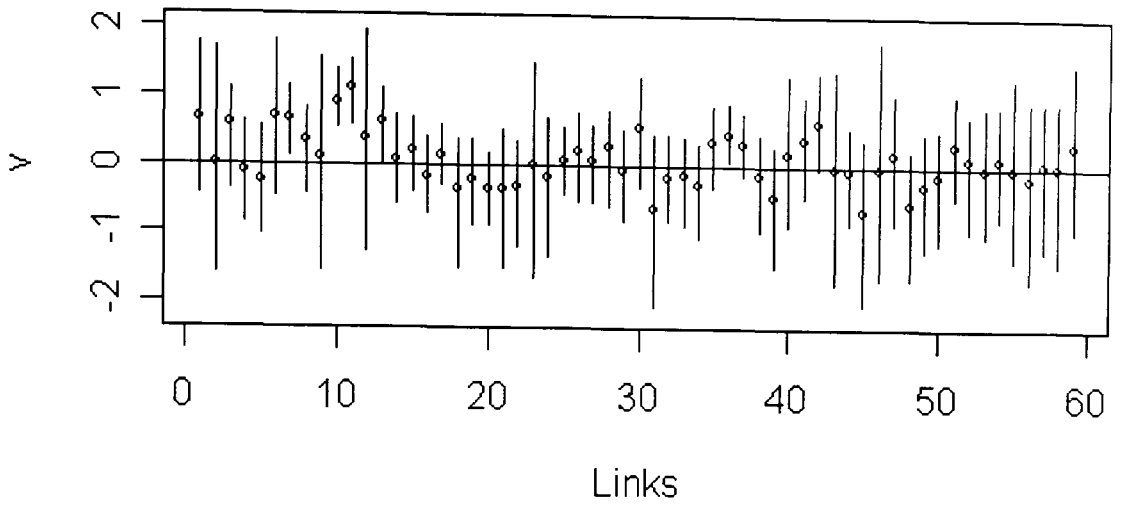

Figure 7.7: 95% credible intervals of log-normal random effects (v) from model PLN.

residuals based on the true y, that corresponds to function $I(y, \lambda^{(j)})$ in Section 4.7. It suggests that spatial correlation in Bayesian residuals is larger than 0.4 and significant. This result is consistent with the result in Table 7.1. By using predicted values for the response variable y, histogram (b) shows that most values of Moran's I ($I(y^{(j)}, \lambda^{(j)})$) fall within the interval between -2 and 2 and are nonsignificant. Therefore, the p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$ is 0. It indicates a misfit of the model.

Significant spatial correlation in residuals is identified for model PL from the above analysis. Moreover, results from model PLN suggest that the inclusion of unstructured random effects has explained a level of spatial variation in the response variable and residuals from such a model are not spatially correlated. Another approach for taking account of spatial variation is to include spatially structured random effects. In order to compare the power of unstructured random effects and spatially structured random effects to explain the spatial variation, an intrinsic CAR model, that includes only spatially structured random effects, was fitted. Figure 7.9 shows values of Moran's I on residuals from this model. Again, both true values and predicted values of y were used to calculate residuals. Both histograms suggest that spatial correlation in Bayesian residuals is nonsignificant. The two-sided p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$ is 0.90. It indicates that the fit of model

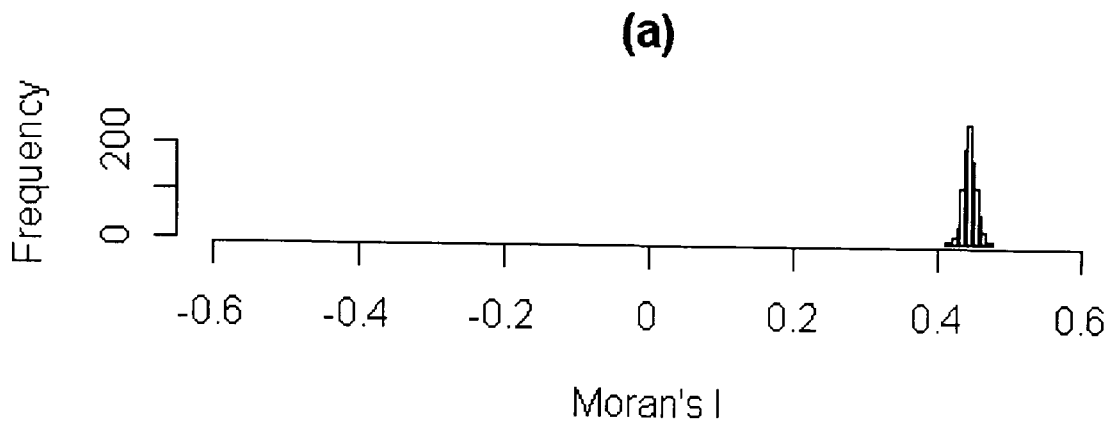

Figure 7.8: Values of Moran's I in Bayesian residuals from model PL: (a) based on true y; (b) based on predicted values for y.

is fine. Figure 7.10 shows the 95% credible intervals of the spatial effects in the intrinsic CAR model. Compared with Figure 7.7, there are more links with positive effects. Moreover, several links are found to have negative effects. The DIC of the intrinsic CAR model is 2307, which is approximately 100 smaller than the DIC of model PLN. This suggests that for the M1 data, models with spatially structured random effects perform better than model with unstructured random effects.

The contribution of the spatially structured random effects and unstructured random effects on explaining extra spatial variation was examined individually. When both of them are included in the model, their relative strengths can be compared. Figures 7.11 and 7.12 illustrate the 95% credible intervals of spatially structured random effects and of unstructured random effects in model CCARtr. The estimates of spatially structured random effects in the convolution CAR model (in Figure 7.11) look similar to those in

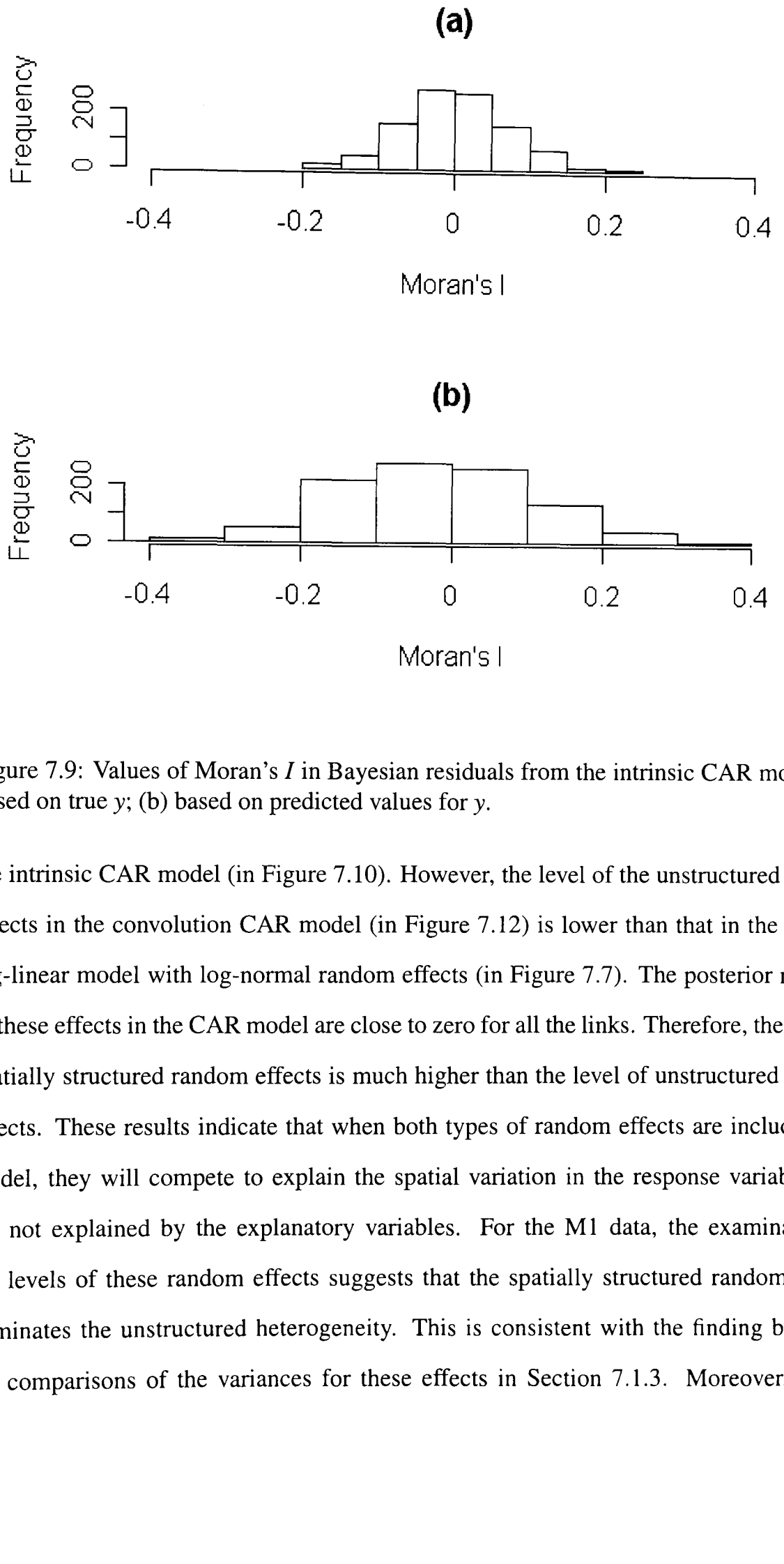

Figure 7.9: Values of Moran's I in Bayesian residuals from the intrinsic CAR model: (a) based on true y; (b) based on predicted values for y.

the intrinsic CAR model (in Figure 7.10). However, the level of the unstructured random effects in the convolution CAR model (in Figure 7.12) is lower than that in the Poisson log-linear model with log-normal random effects (in Figure 7.7). The posterior medians of these effects in the CAR model are close to zero for all the links. Therefore, the level of spatially structured random effects is much higher than the level of unstructured random effects. These results indicate that when both types of random effects are included in a model, they will compete to explain the spatial variation in the response variable, that are not explained by the explanatory variables. For the M1 data, the examination of the levels of these random effects suggests that the spatially structured random effects dominates the unstructured heterogeneity. This is consistent with the finding based on the comparisons of the variances for these effects in Section 7.1.3. Moreover, values

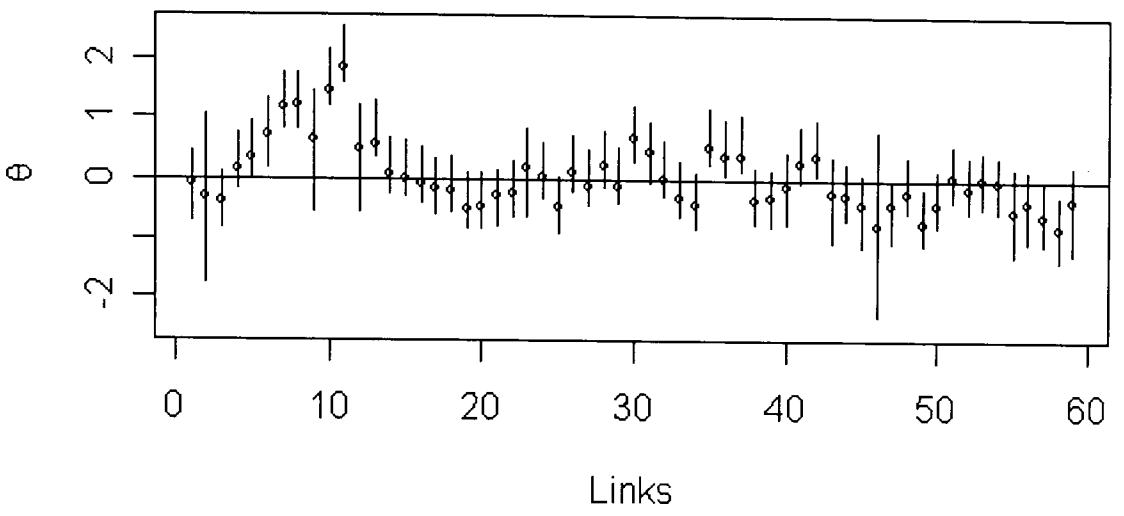

Figure 7.10: 95% credible intervals of spatially structured random effects (θ) from the intrinsic CAR model.

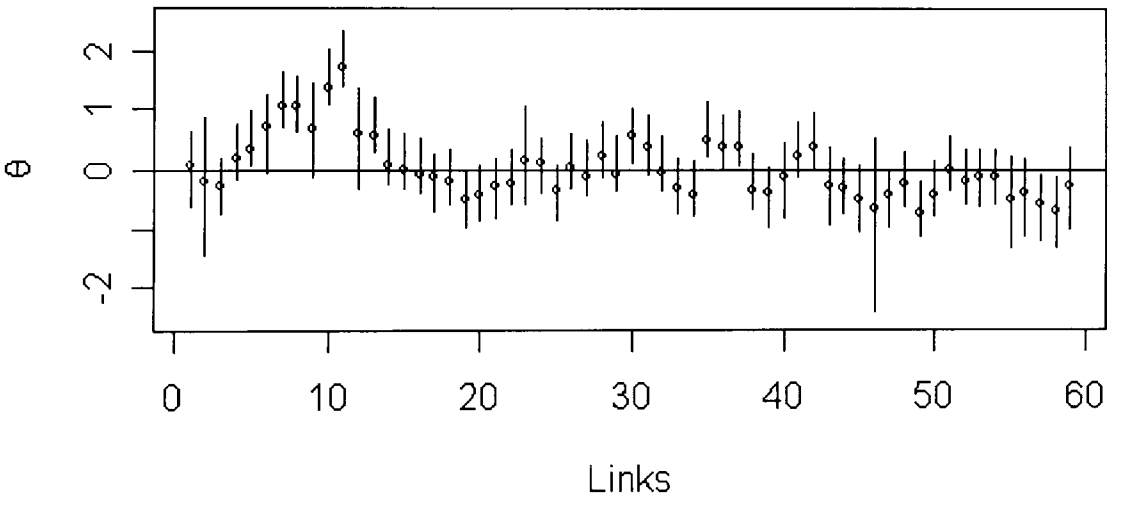

Figure 7.11: 95% credible intervals of spatially structured random effects (θ) from model CCARtr.

of Moran's I for residuals based on the true values and predicted values of y were also obtained and compared. The histograms are similar to those in Figure 7.9 for the intrinsic CAR model therefore are not presented here. The two-sided p-value for $I(y^{(j)}, \lambda^{(j)}) > I(y, \lambda^{(j)})$ is 0.92, therefore suggests a proper fit of the model.

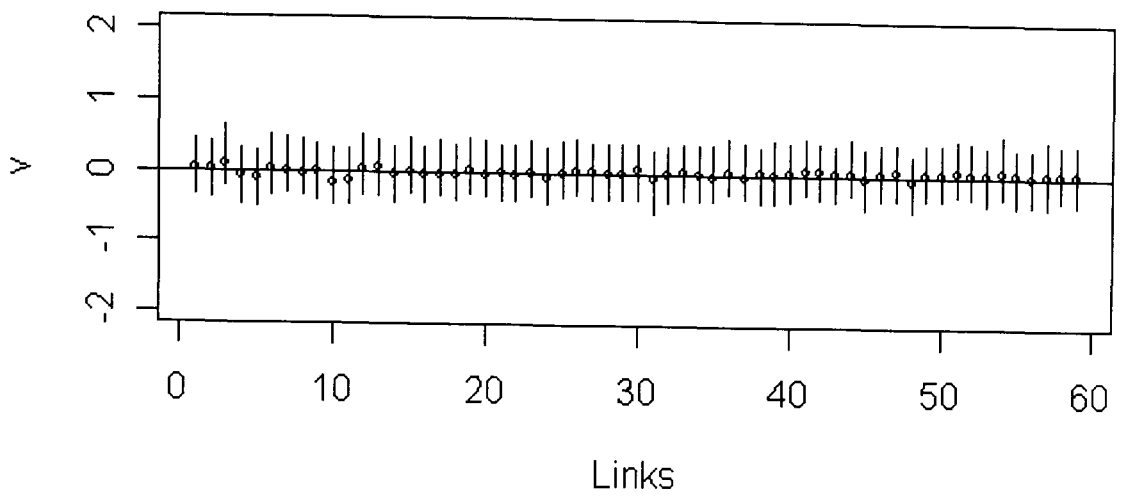

Figure 7.12: 95% credible intervals of unstructured random effects (v) from model CCARtr.

The above analysis and discussion are based on residuals in year 1999. Results using residuals in other years are very similar therefore are not presented here.

7.2 Junction accidents in Coventry

The study network consists of A- and B-roads in Coventry, with junctions the intersections of these roads (defined as major junctions in this thesis). Neighbouring junctions are defined as junctions joined by a common road link. Using the 1-0 weighting scheme, the value of Moran's I in the accident data is about 0.04 which is very small. When the spatial weights are determined by the shortest road sections between neighbouring junctions, the value of Moran's I is marginally significant. It has a value of 0.173 with a p-value of 0.06. These results suggest that very little spatial correlation is in the accident data.

Figure 7.13 shows the distribution of the accident data grouped by junction types. It suggests that roundabouts are likely to have more accidents than other types of junctions. However, previous studies find that roundabouts are usually associated with less accidents than other types of junction. Result here could be just a special case.

Four models were developed for junction accidents. The first three models do not

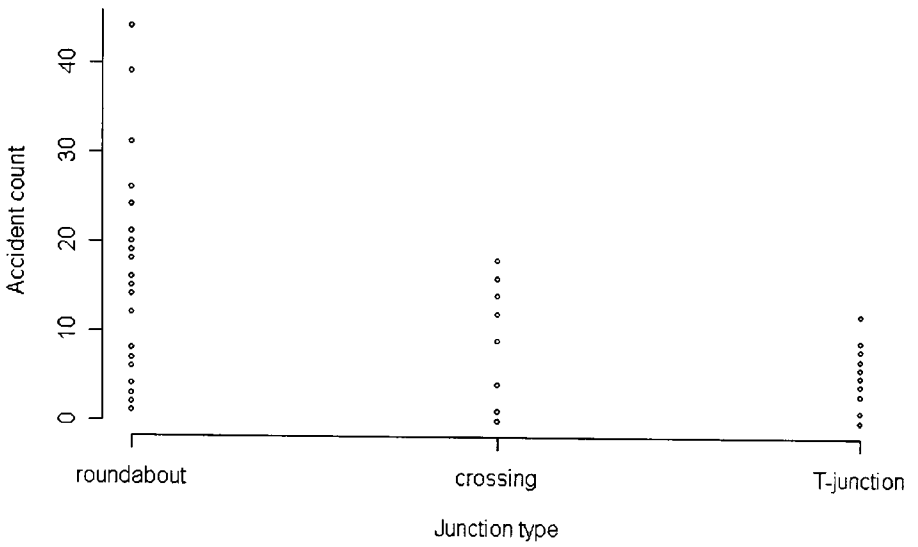

Figure 7.13: Accident counts by junction types

include the explanatory variable junction type but include a constant term and random effects only. They are namely the Poisson model with log-normal random effects (PLN), the CAR model (CCAR) and the proper CAR model (PCAR) (specified as expression (3.2) in Section 3.1.2). For the two CAR models, the spatial weights depending on the distance between the neighbouring junctions were used. The spatial correlation in the standardized residuals from model PLN is small and not significant. The DICs of the three models are all around 317. It indicates that the inclusion of spatial random effects does not improve the model. This result has been expected in advance because the spatial correlation in the accident data is small. In model CCAR, the variance parameter of the unstructured random effects is 0.15 and is much larger than that for the spatially structured random effects, which is 0.02 . This indicates that the unstructured heterogeneity strongly dominates the spatially structured heterogeneity. The correlation parameter ρ in model PCAR, which measures the strength of spatial correlation is estimated to be 0.26. This is another indication of the small spatial correlation in this dataset.

The last model includes the junction type as the explanatory variable in model PLN. Two dummy variables were constructed representing three types of junctions, namely roundabout, crossing and T-junction. The DIC is still around 317. The medians of the coefficients for roundabouts and crossings are 1.12 and 0.62. This indicates more accidents

tend to occur at roundabouts and crossing than at T-junctions.

7.3 Conclusion

In this chapter, the analysis of M1 link accidents and Coventry junction accidents shows very different results. Measured by Moran's I statistic, the spatial correlation in link accidents on the M1 is very high even after controlling for the level of traffic. But the junction accidents in Coventry do not display much spatial correlation. One possible reason for the small spatial correlation is that only major junctions were selected in this study and they may be too far away from each other. The extent of spatial correlation between two major junctions are weak because of the existence of several minor junctions and links between them. Therefore, a study that includes all the junctions in Coventry may give a very different result.

Results show that for M1 link accidents the inclusion of a CAR prior improves the DIC, and successfully removes the high spatial correlation in the residuals. Since both the traffic level and the link length have been considered in the M1 model, the spatial random effects modelled by the CAR prior are likely to capture the similarity in the road characteristics (for instance, the curvature and the gradient) of the neighbouring sites. Moreover, the Bayesian approach to examining spatial correlation in the residuals, introduced in Section 4.7, was used for the M1 models. Results suggest that different results for Moran's I can be obtained for the model with log-normal random effects when using different types of residuals – Bayesian residuals and residuals based on the posterior mean of λ. For Coventry junction accidents, the CAR models does not perform better than the non-CAR models.

Chapter 8

Applications of the models

Areal models and models for road accidents on a road network have been developed in this research. Results from these models show the importance and the influence of the inclusion of the spatial and the temporal effects. In most cases, the inclusion of the spatially structured random effects modelled by a CAR prior improves the performance of the models. When such spatial effects are not included in the models, significant spatial correlation in the residuals has been identified for models using different datasets. The main use of CAR models is to take account of the contributory factors (for the response variables) that are unknown or unmeasurable but are spatially correlated. The inclusion of the temporal effects can successfully remove the highly positive temporal correlation in the residuals when longitudinal data are used to fit the models.

In this chapter, some possible applications of the models developed in the previous chapters are suggested. General approaches for using these models will be explained. Models developed in this research can be used to predict the expected number of accidents in the future. Moreover, the spatially structured random effects can be estimated. Based on them, local authorities or road sites (links or junctions) can be ranked.

8.1 Prediction of accident counts

This section aims to show how spatial models developed in this thesis can be used to predict numbers of accidents in the future. Data for traffic and population in 2006 were

obtained from Department for Transport (Department for Transport, 2007*d*) and National Statistics (Office for National Statistics, 2007). Based on the estimates of parameters in areal models fitted by data from 2001 to 2005 (see Section 6.3), numbers of fatal and serious accidents at the local authority level in 2006 are predicted. Since STATS19 data in 2006 (Department for Transport, 2007*b*) are available, comparisons between the predicted number of accidents and the observed number in each local authority can be made.

In an areal model, the response variable is the number of a particular type of accidents in a local authority in one year. In order to estimate or make a prediction of the expected number of accidents in a year, several conditions must be guaranteed. Firstly, the explanatory variables in the year for prediction need to be available. Secondly, the neighbours list should be unchanged. This means that for areal models the boundaries should not be changed in the year for prediction. Thirdly, the random effects, such as the spatially structured random effects, the unstructured random effects and the temporal effects, need to be predicted. This is straightforward only if spatial effects are constant over time. However, results from areal models developed in this research shows that, when the spatial effects over time are assumed to be constant, models do not fit well. The parameters do not converge. This introduces difficulty for predicting the number of accidents at the local authority level. As shown in Section 6.3.3, the distribution of and the level of spatial effects in an England map is similar in the first three years (2001-2003) as well as in the last two years (2004-2005). Therefore, a compromise approach is to divide the time in two periods and to assume that the spatial effects are constant in each period and assume spatial effects in 2006 are same as in 2004-2005.

Prediction using statistical models in a Bayesian framework can be treated as a missing data problem (see Gelman et al., 2004, Section 21). In the application of predicting accident counts in 2006 using areal models, data for explanatory variables from 2001 to 2006 and accident data from 2001 to 2005 were used to fit the models. Data for the response variable y (number of fatal and serious accidents) in 2006 were treated as missing values. Two models were used to make predictions here. In addition to the explanantory variables, the non-CAR model includes a linear time trend, a dummy variable to iden-

tify unitary authorities, first order autoregressive temporal effects and log-normal random effects. The CAR model extends the above model by including a CAR prior to capture spatial random effects. For the reason discussed earlier, in this CAR model, the extent of spatial effect in each local authority is assumed to be constant from 2001 to 2003 as well as from 2004 to 2005. Results from predictions are discussed below.

Figure 8.1 compares results from predictions using two models based on the posterior medians of λs, the Poisson means. The solid line in each graph is a 45 degree line. If there is no bias in the predictions, equivalent numbers of points should lie over and below the 45 degree line. The figure suggests that, by using a non-CAR model, there are many more points lying over the line than those below the line. By using a CAR model, although there are still more points lying over the line, the problem is not as serious as that for the non-CAR model.

However, in a Bayesian context, λ is random and has a posterior distribution. Therefore, prediction of y in a local authority should be made based on the posterior predicted values of λ. In this application, for each local authority, 1000 simulations of λ in 2006 were saved. For each simulated value of λ, a y^* was simulated from Pois(λ). For each local authority, 1000 predicted values of y^* were obtained. They formed a posterior predictive distribution of y^*. This was compared with the true value of y according to STATS19 data in 2006.

Figures 8.2 to 8.5 show the results from predictions using the CAR model for London boroughs, metropolitan districts, unitary authorities and other local authorities. In each figure, local authorities are ordered by the posterior median of y^*. Here, the 95% credible interval (CI) of y^* is used to summarize the predictive distribution of y^*, illustrated by a horizontal line in the figures. The true y, number of fatal and serious accidents in 2006, is represented by a point. Figure 8.2 shows that the true y is significantly under-estimated for the City of London and for Westminster. Figure 8.3 suggests that the true y is contained by the 95% credible interval of y^* for all the metropolitan districts. Figure 8.4 shows prediction results for unitary authorities. It suggests that the true y is significantly under-estimated for the following unitary authorities: Bracknell Forest, Milton Keynes, York,

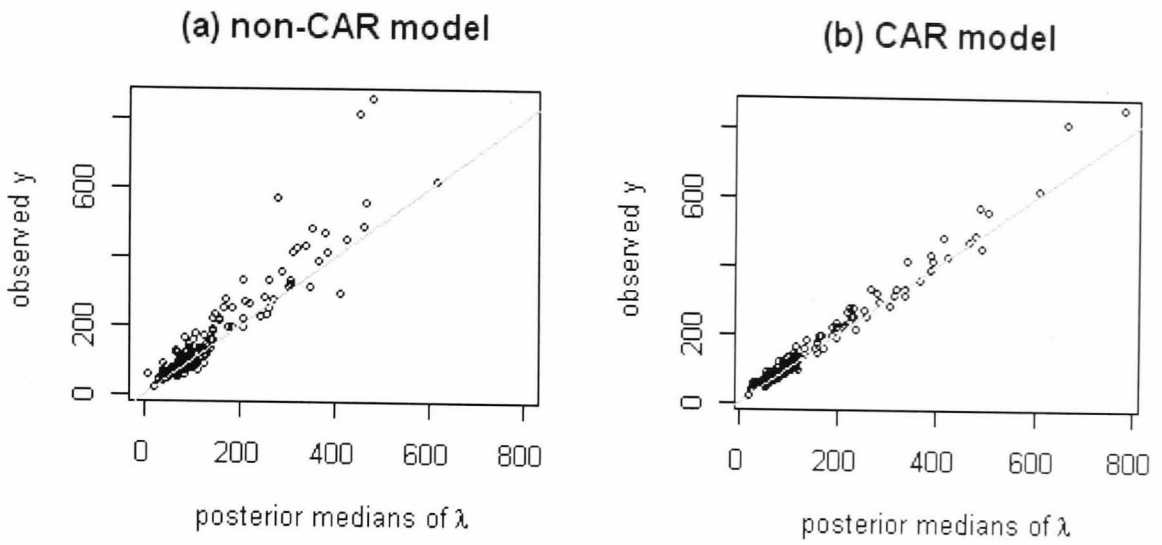

Figure 8.1: Comparisons of prediction results from the non-CAR model and the CAR model by using the posterior median of λ.

Darlington, Blackpool, Brighton and Hove, and Redcar and Cleveland. Prediction results for other local authorities are illustrated in Figure 8.5. According to the figure, the true y is significantly under-estimated for East Sussex, West Sussex and Lancashire.

Results from predictions suggest that most significantly under-estimated y^*s are for unitary authorities. As shown in Section 6.3.5, unitary authorities were found to have more fatal and serious accidents than other local authorities. Here, in the prediction models, the coefficient for the dummy variable that identifies unitary authorities is also positive. This means that the effect of unitary authorities has been taken account of in the model. Therefore, reasons for significantly under-estimating number of fatal and serious accidents in some unitary authorities may need further investigation.

Figure 8.6 illustrates the trend of fatal and serious accidents in the unitary authorities where y is significantly under-estimated in 2006. It shows that there is a big jump in 2006 from 2005 in these unitary authorities. Traffic volumes and population in these areas were found to slightly increase or be stable in 2006 compared with 2005 except for Blackpool unitary authority where the traffic volume slightly decreased in 2006. If these variables were measured correctly, the result will suggest that the large increase in number of accidents cannot be fully explained by the changes in traffic and population. Although as a Poisson distributed variable, y has some extent of random variation, the variation for

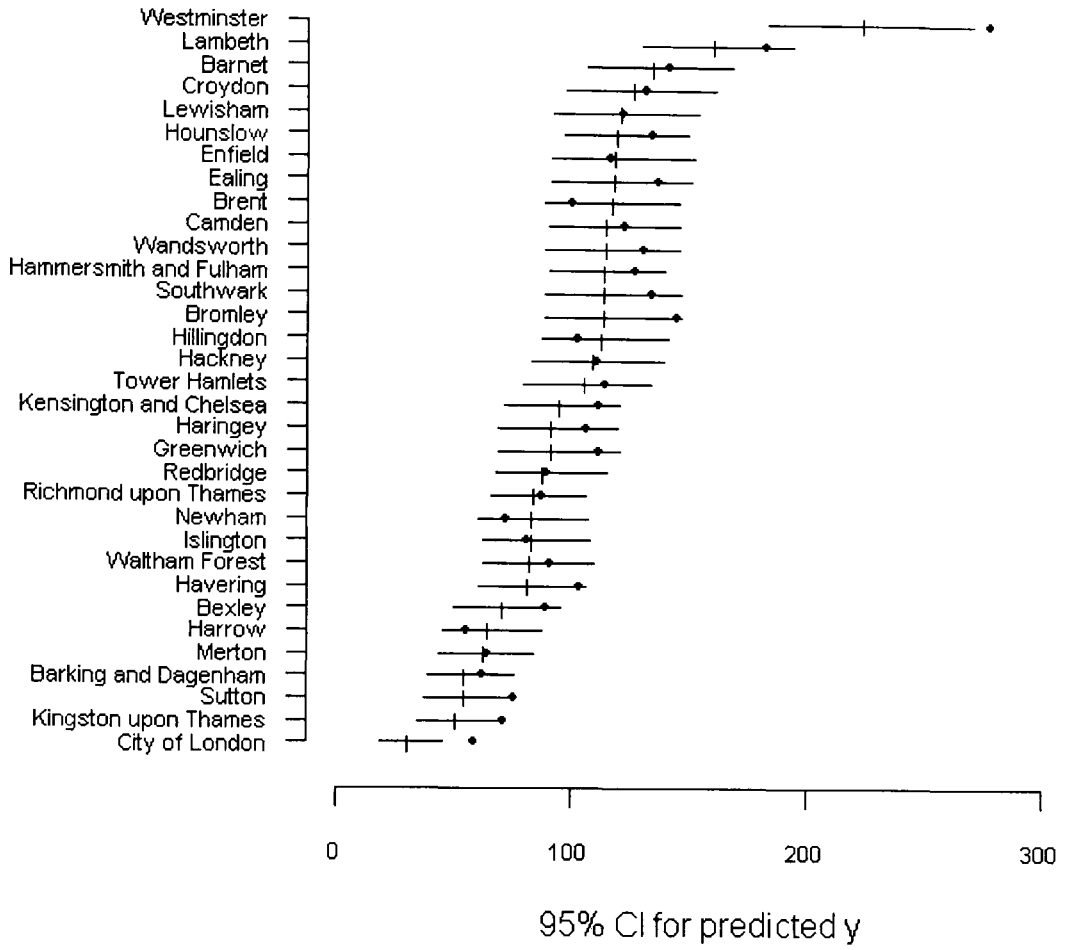

Figure 8.2: Predictions for London boroughs.

the seven unitary authorities shown in Figure 8.6 seems too large.

Figures showing prediction results from the non-CAR model are included in Appendix E. The main difference between the prediction results using different models is that the variation in y^* (predictive values for y) is much smaller from the CAR model than that from the non-CAR model. This suggests that the estimates are more precise from the CAR model. The smaller variation in y^* from the CAR model is due to the smaller variation in the λ, from which y^* was simulated ($y^* \sim \text{Pois}(\lambda)$). In a CAR model, the estimate of λ in an area depends on not only the coefficients of explanatory variables but also the spatial random effect in the area that shrinks towards a local mean depending on its neighbours. This shrinkage may be a possible reason to explain the smaller variation in the estimates of λ from a CAR model.

According to the comparisons of prediction results using different models based on the

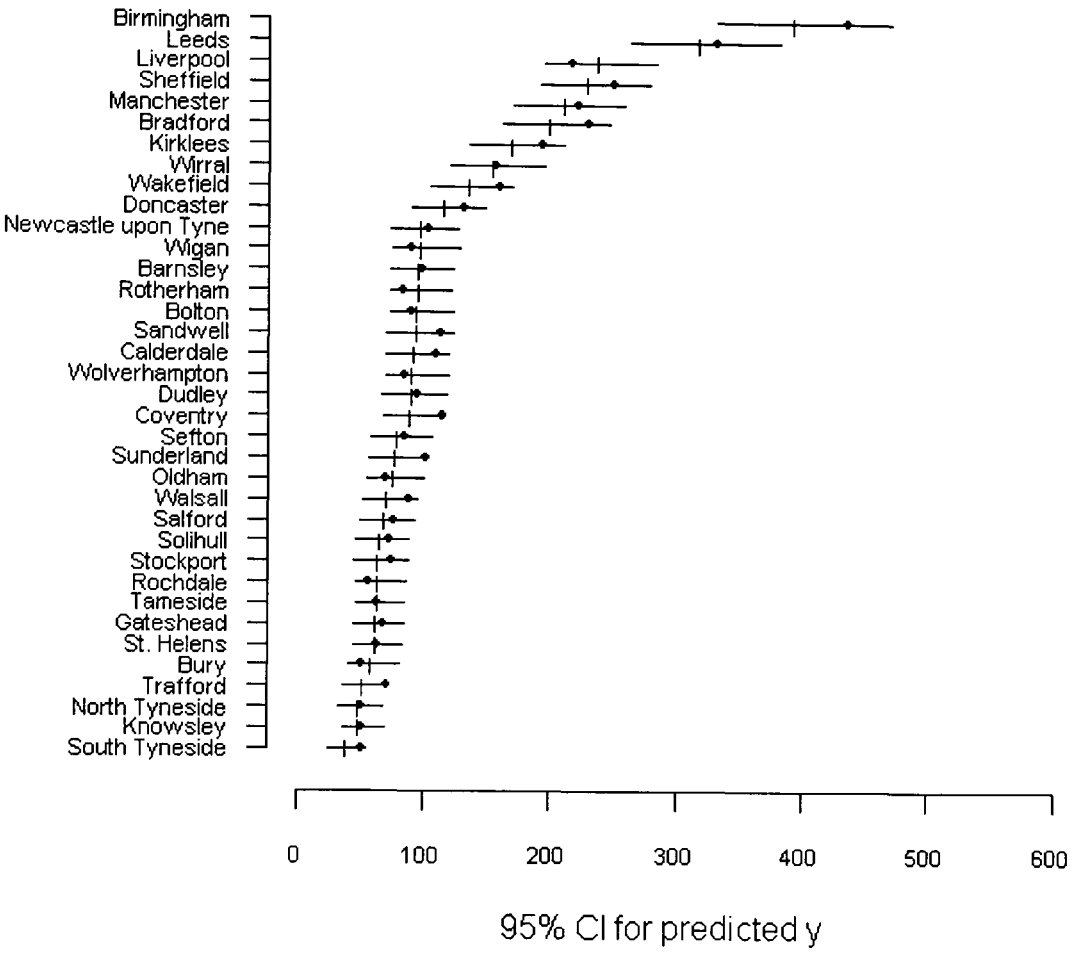

Figure 8.3: Predictions for metropolitan districts.

predictive distribution of y^*, which model is better for prediction is hard to say. If considering local authorities in which accident numbers are significantly under-estimated, seven local authorities were identified using the non-CAR model while twelve local authorities were identified using the CAR model. Therefore, there are fewer local authorities in which accident numbers are significantly under-estimated using the non-CAR model although the difference is small (only five). However, as shown earlier, larger variation in the predicted λ was found for the non-CAR model. Therefore, the posterior predictive distribution of y^* for each local authority obtained from the non-CAR model contains more values and has a wider range compared with that obtained from the CAR model. This possibly explains why there are fewer number of local authorities being significantly under-estimated by using the non-CAR model. If the prediction for the number of accidents in a local authority significantly under-estimates the true value, further research may

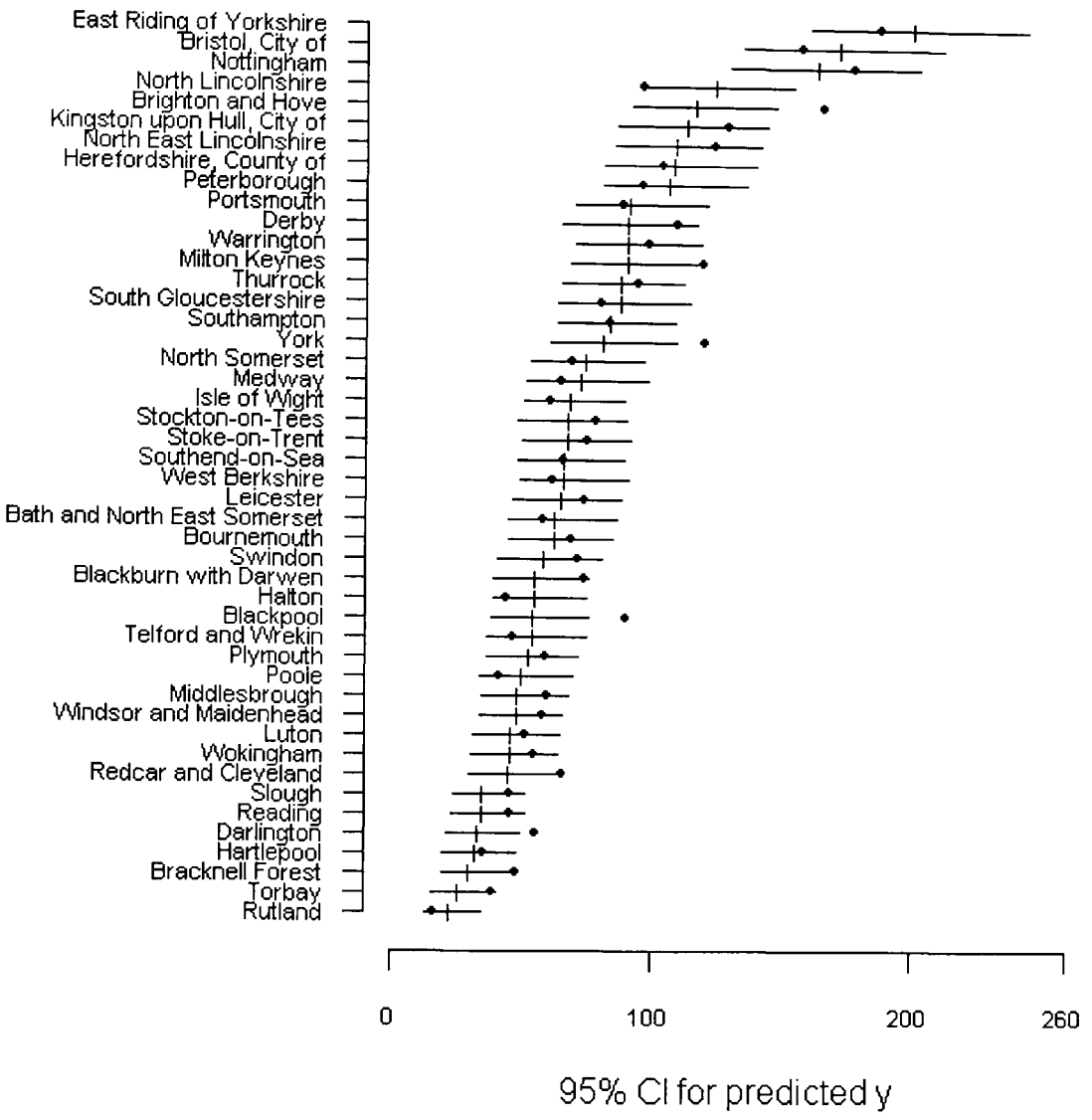

Figure 8.4: Predictions for unitary authorities.

be needed to find out possible reasons. One advantage of using the CAR model in this application is that it produces more precise predictions than the non-CAR model. That is the variation in predictive values y^* is much smaller from the CAR model than those from the non-CAR model. If the predictive distribution of y^* has a very wide spread, as obtained for the non-CAR model, the prediction may not be very helpful in practice. From this view, the CAR model is better for prediction.

The above method for accident prediction is explained using an example of areal models. The same method can be used for predicting number of accidents on a road network, for instance, using the M1 models developed in the previous chapter. However, the traffic flow data for the M1 are not currently available for a new year. Therefore, this cannot be

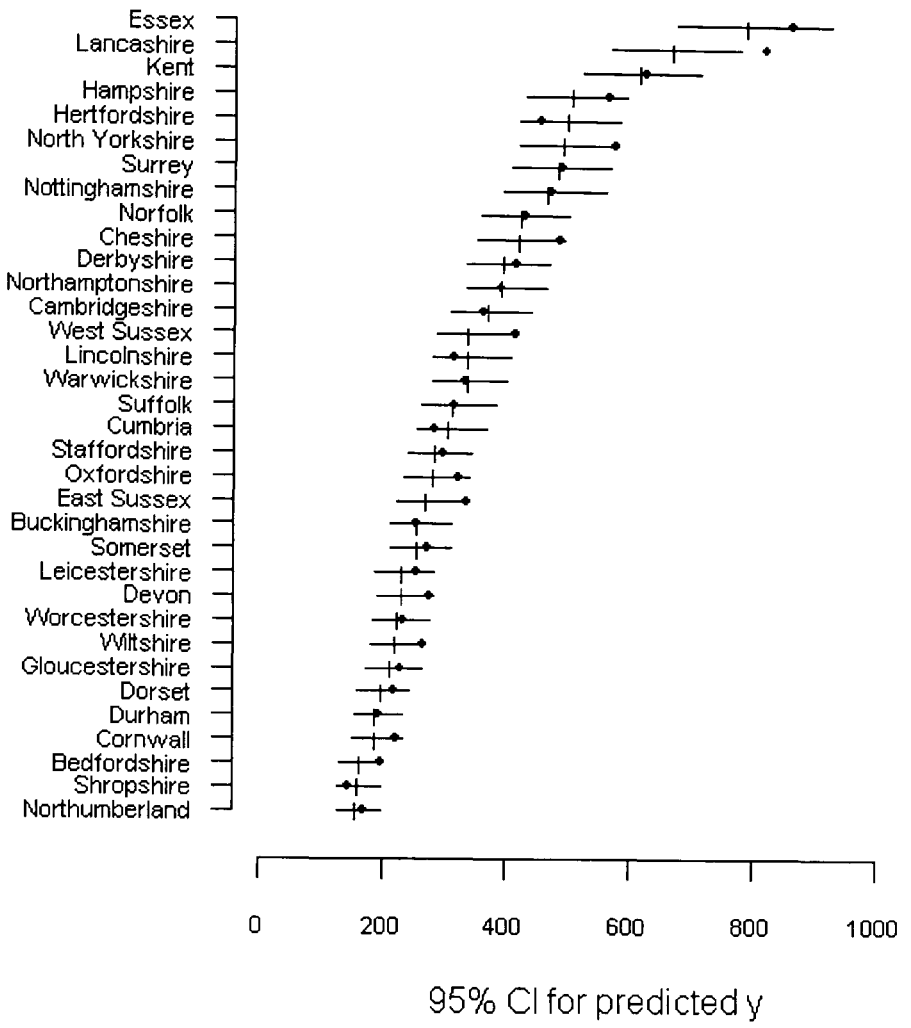

Figure 8.5: Predictions for other local authorities.

achieved in this research.

8.2 Ranking the sites

8.2.1 Background

There is a long history of site ranking in the context of road safety research. The main aims of site ranking are to identify sites with high accident risks and select sites for engineering treatment. There are a number of methods to rank the sites. Most of them are based on the measurements of the accident risk include raw accident rates (observed accident count per vehicle-km) and model-based accident rates (accident means divided by the

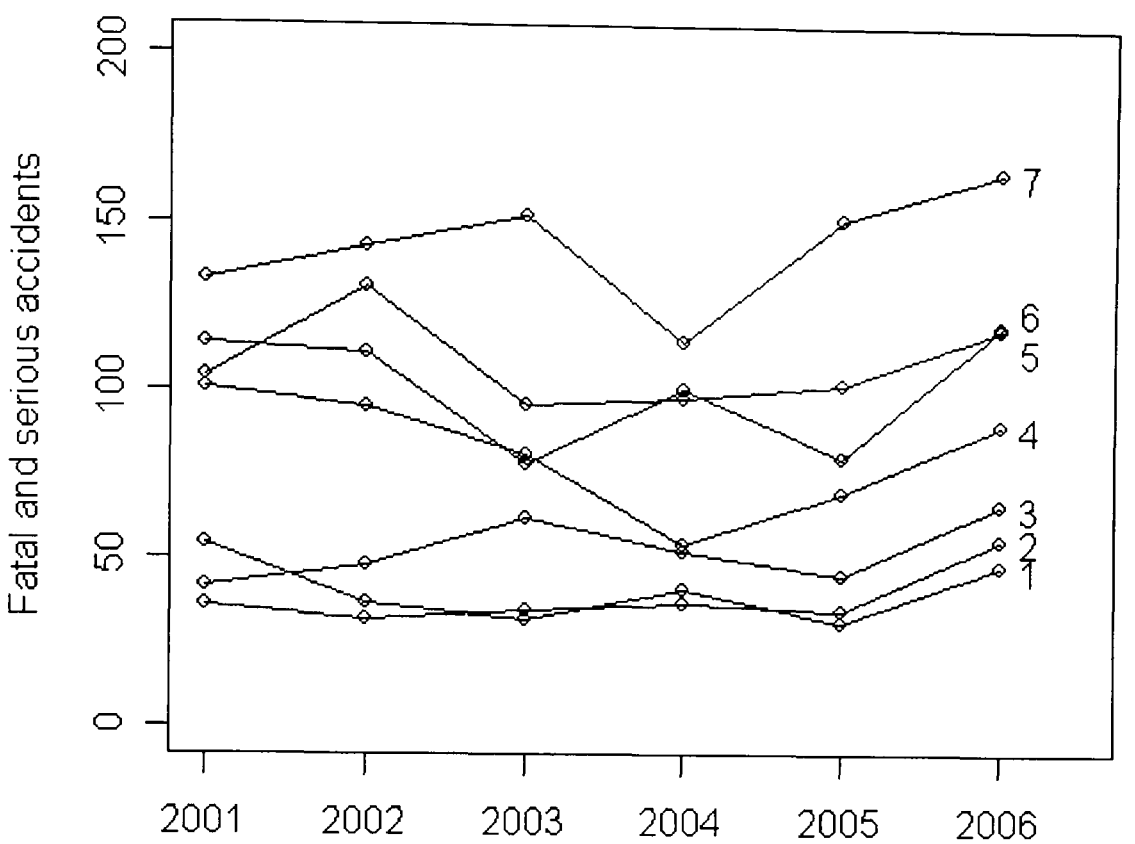

Figure 8.6: Trend of fatal and serious accidents in unitary authorities where y is significantly under-estimated: 1. Bracknell Forest; 2. Darlington; 3. Redcar and Cleveland; 4. Blackpool; 5. Milton Keynes; 6. York; 7. Brighton and Hove.

traffic volume) (Hauer et al., 2004, see). Model-based ranking studies in road safety often use the empirical Bayes estimates of the accident means to compare sites (for instance, Hauer et al., 2004, 2002). Some researchers have adopted the full Bayesian method to estimate the accident means and make comparisons of sites (for instance, Miaou and Song, 2005; Tunaru, 2002). Miaou and Song (2005) discussed the advantages of using the estimates based on the full Bayesian models. The full Bayes estimates take full account of the uncertainty associated with the estimates of the parameters and can provide exact measures of the uncertainty. Three criteria have been much used to rank sites in the literature. They are:

- Ranking by the posterior means or medians

- Ranking by the posterior distributions of ranks
- Ranking by the probability that a site is the worst

8.2.2 Model-based ranking

Models that have been developed in this research take account of the spatial correlation in the accident means. These models have been found to produce better estimates of the accident means therefore ranks based on these models can be more reliable. The straightforward method to rank the sites is to rank the estimated accident means $\hat{\lambda}$ of all the sites. But to compare the risks of the sites, accident rates $\hat{r}$ based on the $\hat{\lambda}$ are more appropriate to be used. When longitudinal data are used, the mean accident rates need to be calculated. Consider the M1 data used to fit the link models in Chapter 7, the mean accident rate at link i can be expressed as:

$$\hat{r}_i = \frac{\sum_{t=1}^{T} \hat{\lambda}_{it}}{365 \times LENGTH_i \sum_{t=1}^{T} AADF_{it}}, \tag{8.1}$$

where $AADF_{it}$ is the traffic flow in link i in year t and $LENGTH_i$ is the length of link i. Ranks based on the mean or the median of $\hat{r}_i$ can be obtained. However, ranks are uncertain and the uncertainty associated with them are important to be examined. Therefore, several researchers have adopted the posterior distribution of the ranks based on the posterior distribution of λ (for instance, Miaou and Song, 2005; Tunaru, 2002).

In addition to rank the sites based on the accident rates, the spatial models developed in this research provide an additional measure to compare the sites. The spatial random effects, which are used to take account of the spatial correlation in the accident means at different sites, can be estimated from the models. They measure the amount of the spatial effects that are spatially correlated but have not been explained by the explanatory variables. Therefore, ranks by the random spatial effects can be used to identify sites with high risk caused by some unobserved factors which are spatially correlated. The worst sites selected by such a criterion need special attention because factors that influence the safety of the sites are not clear. Further on-site investigation may be needed to find out

the causes of high accident risk.

By adopting the above mentioned methods to rank the sites, links on the M1 have been ranked based on the estimates from the link models developed in Chapter 7. The raw accident rates, accident rates based on a non-CAR model and accident rates on a CAR model have been obtained using equation 8.1. Figures 8.7 and 8.8 show the difference in ranks using different measurements of the accident rates. Figure 8.7 plots the ranks based on the estimates from a CAR model (model CCARtr in Table 7.1 in Chapter 7) against the ranks by the raw accident rates. Figure 8.8 plots the ranks based on the CAR model against the ranks based on the estimates from a non-CAR model (model PLNtr&re in Table 7.1). The figures show that the ranking results can be different when different types of accident rates are used. For instance, in Figure 8.7, the two points lying furthest away from the 45 degree line correspond to the two spurs represented by the two vertical links in the node-link graph in Figure 5.2 in Chapter 5. Using the raw accident rates, their ranks are higher than 20. However, they are ranked 7th (the left one) and 6th (the right one) based on the CAR models and 12th and 18th based on the non-CAR models.

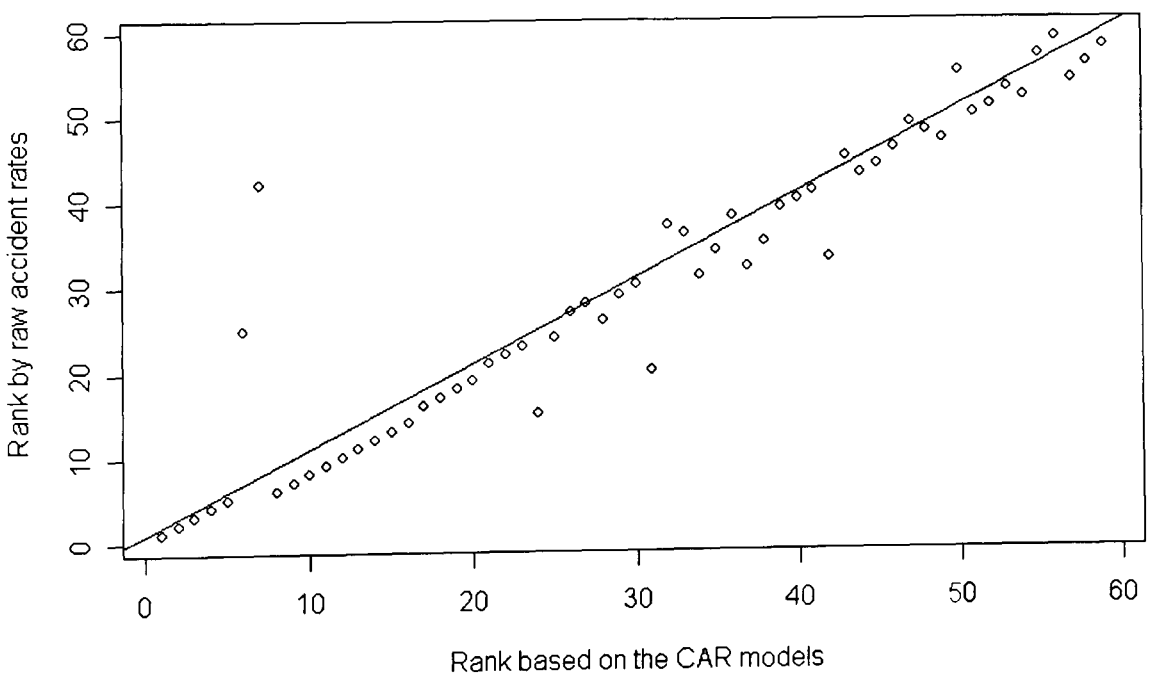

Figure 8.7: Comparisons of ranking results: A.

Figure 8.9 illustrates the estimated ranks of the accident rates based on the CAR

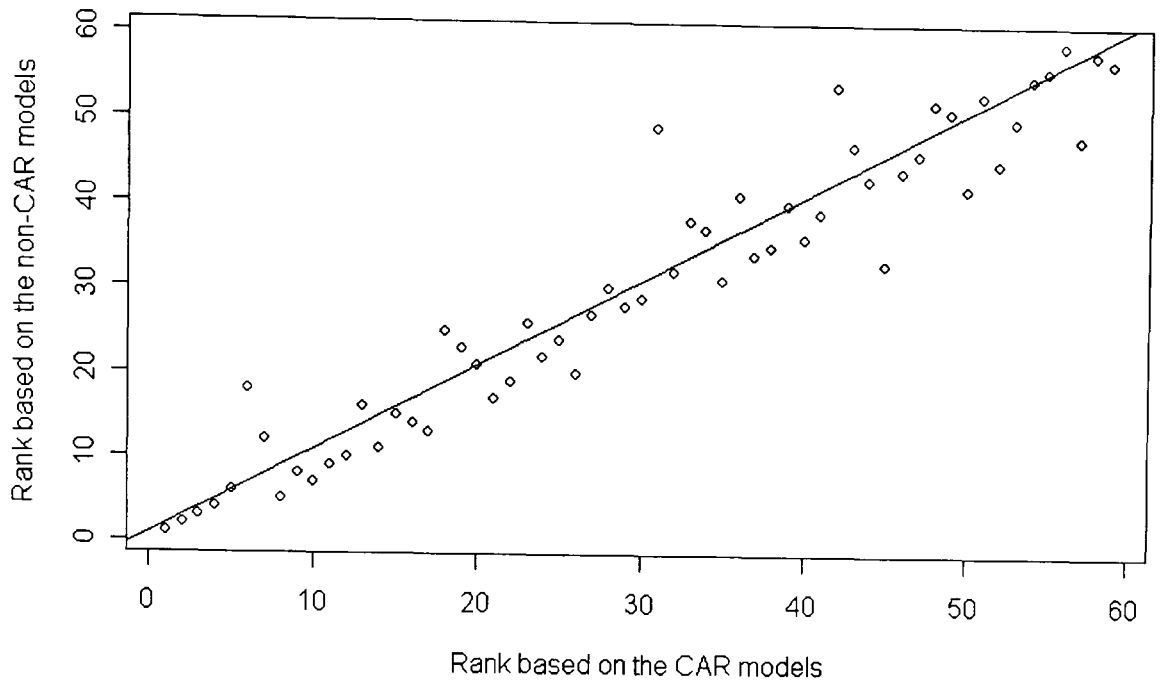

Figure 8.8: Comparisons of ranking results: B.

model. The last 1000 iterations on the λ when fitting the models have been saved and used to produce this plot. For each link on the M1, the median of its rank (plotted by a dot) and the lowest and the highest rank (represented by edges) are shown in the plot. According to the plot, links 7, 8, 10 and 11 are identified as the worst links. They are neighbouring road links between junction 6A and Junction 10 on the M1 in Hertfordshire and a very small part of Bedfordshire that intersect with the M25, M10, A4147 and A5183. Other links whose ranks are high and have small variation include links 13, 35, 36 and 37. Link 13 is in Bedfordshire. The last three links are road sections between junction 27 and junction 30 on the M1. Link 35 crosses the boundary between Nottinghamshire and Derbyshire. Links 36 and 37 are in Derbyshire.

As introduced earlier, the estimates of the spatial random effects in a CAR model allow for another way to rank the sites. Figure 8.10 illustrates the 95% credible interval of the estimated spatial random effects for each link on the M1. The figure shows that the spatial random effects are strong for links 7, 8, 10 and 11. The medians for them are over 1. Because the spatial effect is an additive term linked to the log Poisson mean, the expected number of accidents at a site with spatial random effects estimated to be 1 will

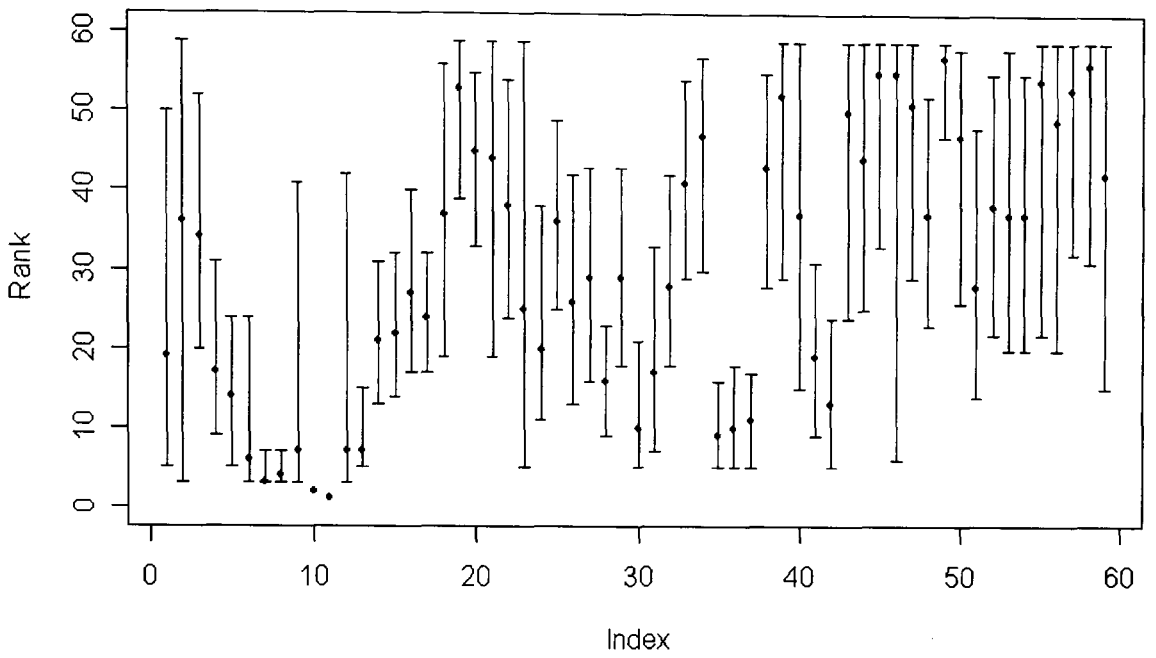

Figure 8.9: Posterior ranks by the accident rates.

be 2.7 times of that at a site with very small random effects (say zero). This suggests that the spatial random effects relating to these links can have great influence on the risk of these links.

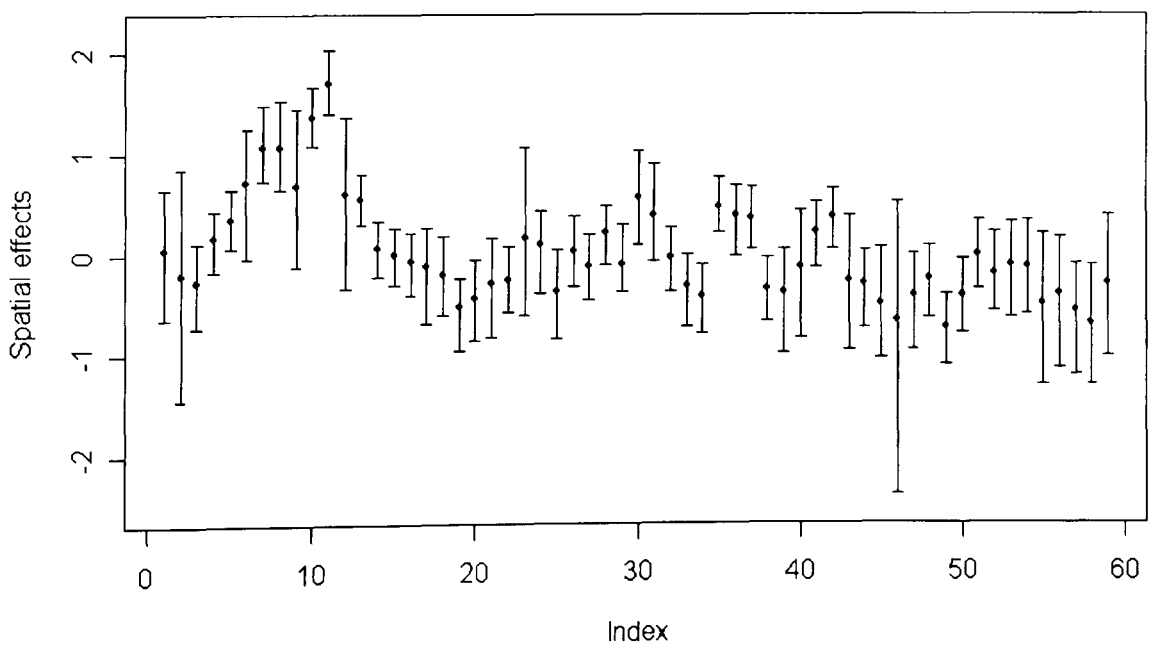

Figure 8.10: Spatial random effects for the M1 links.

Figure 8.11 plots the posterior ranks of the sites by the spatial random effects. For some of sites, there is much variation in their ranks. Links 7, 8, 10 and 11 are again identified as the worst links. Ranking results from the raw accident rates, the model-

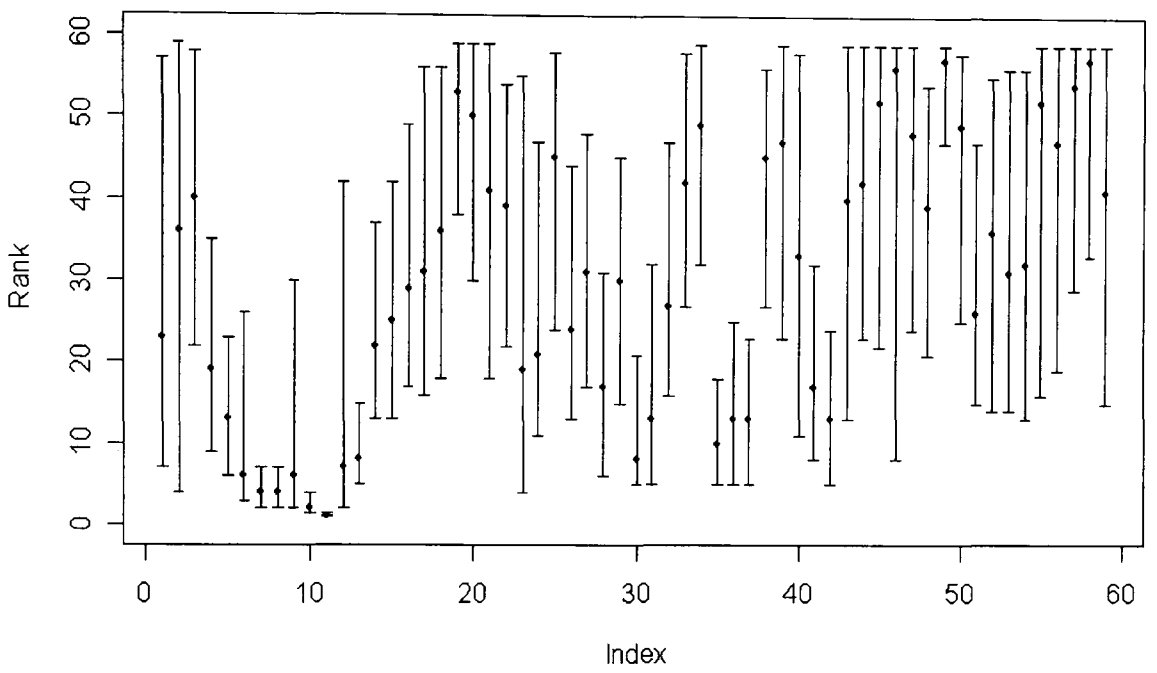

Figure 8.11: Posterior ranks by the spatial random effects for the M1 links.

based accident rates and the spatial random effects show that the links (7, 8, 10 and 11) between junction 6A and junction 10 on the M1 have been ranked in the very top positions by different ranking approches. This suggests that these links need special attention.

In addition to use the estimated random effects to rank sites on a road network, this approach can also be applied to rank local authorities. Recall Figure 6.23 in Chapter 6 which illustrates the 95% credible intervals of the spatial random effects estimated from model CCAR(t)tr.temp in each local authority in 2001. Figure 8.12 illustrates the posterior ranks by the spatial random effects for fatal and serious accidents at the local authority level. The variation in the ranks is large. Local authorities $30-54$ are found to rank in the top positions according to their median ranks. They are actually London boroughs. Names of these boroughs can be found in Table A.2 in Appendix A.

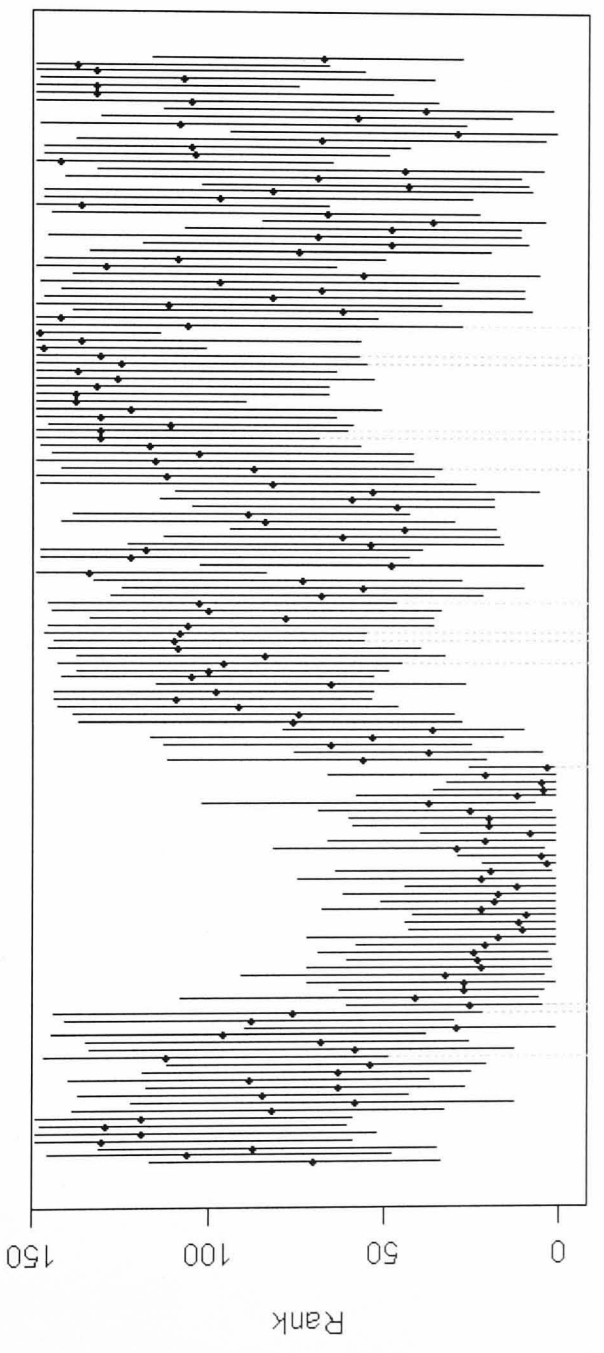

Figure 8.12: Posterior ranks by the spatial random effects in local authorities.

8.2.3 More on the M1 links

The ranking results for the links on the M1 suggest that links 7, 8, 10 and 11 between junctions 6A and junction 10 have high accident risk. If the traffic levels have been measured precisely and included in the models, ranks based on the spatial random effects may indicate that there are other factors that are associated with the high accident rates on these links. These links are neighbouring links. The physical characteristics of the neighbouring links, for instance the curvature or gradient, are more likely to be similar. Moreover, these links experience long delays at peak times. When high speed roads are overloaded, the erratic changes in speed are quite dangerous. Therefore, congestion could also be a relevant factor to explain the high risk identified in these links. Since no variables relating to the road characteristics were included in the M1 model, the unknown contributing factors captured by the spatial random effects are expected to be something relating to the road characteristics. This suggests that road construction or treatment projects on this section of the M1 need special attention regarding to the road safety aspect.

The M1 is a strategic link connecting London, the Midlands and the North. It is one of the busiest motorways with serious delays in some sections. Currently, there are some ongoing widening projects on different sections of the M1. More widening projects in the future for the M1 have been planned (see Highways Agency, 2007). The section between Junctions 6A and 10 on the M1 carries heavy traffic each day with long delays experienced at peak times and high accident rates (see Highways Agency, 2005). Works to widen this section commenced in March 2006 and are expected to be completed by Autumn 2008 (see Highways Agency, 2006). Work will be carried out on both carriageways of the M1 between junctions 6A and 10 to bring it up to a full standard four lane motorway with continuous hard shoulders. The purpose of this project is to reduce congestion and improve both safety and journey time reliability. This widening scheme will be followed with an implementation of the Pilot High Occupancy Vehicle (HOV) lane between junctions 7 and 10 (see ATKINS, 2006). HOV lanes are intended for use solely by vehicles with more than one occupant. They have been used extensively and successfully in the

USA, but only been implemented in the UK on a small number of short sections of dual carriageway (in Leeds and South Gloucestershire). HOV Lanes are aimed at changing travel behaviour to reduce congestion through making better use of the available carriageway. One concern with the HOV lane is its impact on the road safety. Although there are some studies that found the decrease in the accidents after the implementation of the HOV lanes, the report about the M1 HOV lane pilot 'before' monitoring study by ATKINS (2006) cites some previous studies that found the HOV lanes can increase injury accidents. The report suggests that the increase in the accidents can be explained by the speed difference between the two lanes. Therefore the impact of the HOV lanes on the safety needs special attention. The report says 'although the site was not the preferred site in terms of predicted economic return, the programmed widening scheme between J6a and J10 of the M1 would provide the most realistic option for a pilot study in the short term.' It defines the safety objective of the project as follows:

- to have no negative impact on the number and severity of casualties between Junctions 6A and 10;
- the HOV scheme should not cause an increase in risk regarding lane changing accidents on that section when compared to a similar dual 4-lane motorway.

These suggest that when the road sections, especially those with high accident rates in the past, are planned to include HOV lanes, factors associated with the high accident rates need to be investigated before the pilot project and the impact of the HOV lanes on the safety need to be monitored during the project. Moreover, the choice of appropriate road sections to include HOV lanes can be important. From this view, accident prediction models and model-based ranks can be very useful in practice.

Chapter 9

Conclusion

9.1 Summary of the thesis

This research has aimed at developing spatial models for road accidents. These models take account of the spatial autocorrelation between neighbouring areas or sites. When the response variable is the accident count in an area, such as a local authority, the spatial autocorrelation is the correlation between the neighbouring areas, which are often defined as areas that share at least one common boundary. When the response variable is the accident count at a site, such as a junction or a road link, the spatial autocorrelation may exist for neighbouring sites that are usually determined by the layout of the road network. Such autocorrelation can be introduced in a conventional accident model, such as a Poisson regression model with log-normal random effects, by including spatially structured random effects. The conditional autoregressive (CAR) model was used in this research to take account of such spatial effects. Besag, in his contribution to the discussion of McCullagh (2002), suggests the main reason for including spatial effects in a model is to absorb an appropriate level of spatial variation, rather than produce a spatial model with scientifically interpretable parameters. He also views the use of the CAR model in spatial epidemiology as a mainly exploratory approach to account for unknown explanatory variables that are spatially correlated. But if and when such variables are known and are included in the models, a spatial formulation using a CAR prior may not contribute much improvement to the model. In the context of modelling road accidents, there are at least two reasons to

consider spatial effects in the models. First, traffic levels, road geometry and other important variables are difficult to measure perfectly and data for these variables are not always available. When the spatial variation in the explanatory variables is not fully captured by the observed data, the inclusion of spatial effects at least can partly take account of the remaining spatial variation. Secondly, there can be some unknown factors that contribute to accident frequencies. If such factors have similar values in neighbouring areas or at neighgouring sites, models with spatially structured random effects can better explain the variation in the response variables. Both variables like traffic levels and road geometry, and those unknown factors are very likely to be spatially correlated. This is why a CAR model is appropriate to be used in this research.

Both areal models and network models have been developed in this research. The spatial dependency among areas like local authorities and wards can be obtained from the corresponding boundary map while the spatial dependency among sites on a road network is decided by the spatial structure of the road network. Such spatial information is needed for forming a CAR model. Moreover, approaches to including temporal effects in spatial models when data cover two or more periods and jointly modelling different types of accidents have been proposed and examined. Areal models were fitted using data for local authorities in England covering two different periods (one in the 1980s and the other in the 2000s) and data for wards in the West Midlands in one year. Numbers of accidents were disaggregated by severity for all areal models. Network models were fitted to the M1 link data and Coventry junction data. Fitted results from non-CAR models and CAR models were compared and the influence of including spatial effects were examined. They show that adding a spatial CAR component to a conventional accident model has at least one of the following two effects:

- improving the DIC, a measure of the model performance;
- removing the significant spatial correlation in the residuals calculated based on the posterior expectation of λ (the Poisson mean).

Based on these features, CAR models are believed to produce better estimates of the expected number of accidents in an area or at a site. Moreover, the extent of spatial

random effects modelled by a CAR prior can be estimated. These suggest that spatial models developed in this research will be valuable in practice. Two types of applications of such models were introduced in this thesis. The first one is about predicting the number of a particular type of accident in a local authority in a new year. The second one is to use the estimates of spatial effects to rank high-risk sites or local authorities.

9.2 Findings from the analyses

For areal models, results from residual maps show that residuals from non-CAR models are more likely to be spatially clustered and the inclusion of a CAR component may lead to a more random spatial pattern of residuals. CAR models that take account of higher order neighbours and use distance-based spatial weights do not perform better than those considering only first order neighbours and adopting a 1-0 weighting scheme. The DIC for these models are no better than, and sometimes even worse than, the DIC for models that use simpler weights.

When longitudinal data were used, the best performing models were those that assume the spatial effect in an area is not constant over time. This indicates that the unknown or unmeasurable explanatory variables captured by the spatial component may vary over time. Therefore their influences on accident frequencies, measured by the extent of spatial effects, change over time. When spatial effects are assumed to vary over time in a CAR model, the model will not be straightforward to use for predicting accident counts in the future. However, if how these spatial effects change over time could be found and modelled, such a model could be used for prediction.

The variance and estimates of spatial random effects are two ways of measuring the strength of such effects and their influence on accident frequencies. The ratio of the variance of spatially structured random effects against the variance of unstructured random effects measures their relative strength to explain the variation in the response variables. As discussed earlier, the main use of the CAR model is to account for unknown or unmeasurable explanatory variables that are spatially correlated. Therefore, if the strength of spatially structured random effects is found to be fairly strong, it may indicate the exis-

tence of such explanatory variables (unknown or unmeasurable and spatially correlated) and the effect of using the CAR model to take account of them. On the contrary, if the strength of spatially structured random effects is found to be weak, it may indicate that most spatially correlated explanatory variables are known and have already been included in the model. Therefore there is not much left for the CAR model to take account of. For accidents at the local authority level between 1983 and 1986, results suggest that the unstructured heterogeneity dominates the spatial heterogeneity for both serious and slight accidents. For accidents between 2001 and 2005, the spatial heterogeneity and the unstructured heterogeneity have similar strength for fatal and serious accidents but the unstructured heterogeneity dominates the spatial heterogeneity for slight accidents. For ward models, the spatial heterogeneity and the unstructured heterogeneity have similar strength for both types of accidents.

The estimate of the spatial random effect in an area can be summarized by the 95% credible interval of its posterior distribution. Maps of these effects were found useful for identifying areas having positive or negative spatial effects. In local authority models using data from 2001 to 2005, for both types of accidents, districts in metropolitan counties were found more likely to have negative spatial effects except for those in the West Midlands. London boroughs often appeared to have positive spatial effects. Signs of the spatial effects suggest the relative influence of the unknown or unmeasured contributory factors, captured by a CAR prior, on the accident frequencies. An area with a positive spatial effect will have a larger expected number of accidents compared with the number in an area which has a zero or negative spatial effect, but similar values of all the explanatory variables. Similarly, an area with a negative spatial effect will have a smaller expected number of accidents compared with the number in an area which has a zero or positive spatial effect, but similar values of all the explanatory variables. These indicate that conventional models without any spatial effects can over- or under-estimate the accident frequencies.

When correlation between different response variables, representing different types of accidents, was considered in CAR models, the expected deviance became smaller. How-

ever, there were more parameters used in multivariate CAR models therefore the model DIC was not much improved. For models at the local authority level, the within-area correlation between the spatially structured effects for the two types of accidents was found to be fairly high in some years. The within-area correlation between the unstructured random effects was not as high as that between the spatial effects. For ward models, such correlations for the spatially structured effects and the unstructured random effects were similar.

The shared component spatial models that are much used in disease mapping to jointly model two types of diseases were found not suitable for accident data. Normally in models for disease rates there are not many explanatory variables. The shared spatial component for different types of diseases will mostly capture the effects caused by unknown or unmeasurable factors that are spatially correlated. Such effects are shared by both types of diseases. In accident models, such effects are probably mostly captured by the available explanatory variables that are common to both types of accidents. This suggests that very little effect will be left for a shared CAR component to capture.

The estimated coefficient for each explanatory variable in a model can be summarized using the 95% credible interval of its posterior distribution. The sign of a coefficient indicates the form of relationship between the explanatory variable and the expected number of accidents. Using CAR models at the local authority level for serious accidents on built-up A-roads in the 1980s, the number of licensed vehicles, length of built-up A-roads and traffic volumes on built-up A-roads were found to be positively associated with the expected number of accidents while area has a negative association with the number of accidents. For slight accidents, variables other than traffic volume were found not to contribute much to explain the variation in the expected number of accidents. For CAR models using data between 2001 and 2005, the expected numbers of both types of accidents were found to be positively associated with population, length of A-roads, length of B-roads, length of minor roads and number of junctions, but negatively associated with area. Moreover, if two local authorities have the same level of total traffic, the local authority that has a higher proportion of traffic by cars is expected to have fewer accidents

than that in the local authority with a higher proportion of traffic by other vehicles. The estimates of coefficients from a CAR model were found to have larger variation than that in the estimates from a non-CAR model. This is consistent with what Schabenberger and Gotway (2005) suggested – the variability of the coefficients could be under-estimated if autocorrelation in the data is ignored. Moreover, a small and negative linear trend has been identified for both types of accidents. This suggests that accident frequencies at the local authority level decrease over time and models without the trend do not explain this decrease. However, what this trend represents or explains is difficult to identify. For ward models, variables found to be positively associated with both types of accidents are the area, length of major roads, population travelling to work by car as passenger and population travelling to work on foot. The number of junctions is positively associated with fatal and serious accidents but negatively associated with slight accidents. The population travelling to work by bus is positively associated with fatal and serious accidents but has no apparent effect on slight accidents (the 95% credible interval of its coefficient contains both positive and negative values). Moreover, the length of minor roads is negatively associated with slight accidents but has no apparent effect on fatal and serious accidents.

For accidents on a road network, the analysis of link accidents and junction accidents showed very different results. Measured by Moran's I statistic, the spatial correlation in link accidents on the M1 was very high even after controlling for the traffic levels. But the junction accidents in Coventry did not display much spatial correlation. One possible reason for this small spatial correlation is that only major junctions were selected in this study and they may be too far away from each other and therefore the spatial correlation may spread out over the minor junctions and links between them. It may also be possible that the the extent of spatial correlation is different for motorways and urban A- and B- roads. Results show that for motorway link accidents the inclusion of a CAR prior much improved the DIC and successfully removed the positive spatial correlation in the residuals. But there was no apparent influence of spatial effects for Coventry junction accidents. Since both traffic levels and road lengths were included in the link models, the spatial random effects modelled by the CAR prior are most likely

to capture the similarity in the road characteristics (for instance, the curvature and the gradient) of the neighbouring sites.

Models at the local authority level were used to predict the number of fatal and serious accidents in each local authority in 2006, based on previous years' data. One advantage of using the CAR model compared with the non-CAR model for prediction is that it produces more precise predictions than the non-CAR model. That is, the variation in predicted values y^* (predicted numbers of accidents) is much smaller from the CAR model than that from the non-CAR model. The smaller variation in y^* from the CAR model is due to the smaller variation in the λ, from which y^* was simulated ($y^* \sim \text{Pois}(\lambda)$). In a CAR model, the estimate of λ in an area depends on not only the coefficients of explanatory variables but also the spatial random effect in the area that shrinks towards a local mean depending on its neighbours. This shrinkage may be a possible reason to explain the smaller variation in the estimates of λ from a CAR model. By using the CAR model, the predicted numbers of accidents under-estimated the true values in several unitary authorities. An examination of the trend of fatal and serious accidents showed that there was a big jump in 2006 from 2005 in these unitary authorities. Further research may be needed to find out why the observed numbers of accidents are higher than predicted in these areas.

The other main application of the spatial models is a new approach for ranking high-risk sites. This is based on the posterior ranks of the estimates of spatial random effects. Such ranks can identify sites with unmeasured or unknown factors that are associated with higher accident rates. Although such factors are often hidden, at the same time they are very likely to be spatially correlated. Therefore, by observing and comparing the characteristics and the conditions of these sites, further contributing factors are possible to be found. This approach is also suitable to rank local authorities and to identify areas with positive spatial random effects. Further research is needed to investigate reasons for obtaining positive spatial effects in these areas and what they explain.

9.3 Limitations of the research

There are a number of limitations of this research. They can be grouped into three main categories: limitations of the data, limitations of the methods, and limitations for practical use.

The main source of accident data in the UK, the STATS19 database, has some limitations. It contains only accidents involving personal injuries. Therefore, accidents without personal injuries cannot be taken account of in the analysis of accident data. Moreover, STATS19 does not include unreported accidents. This affects the analysis of both accidents and casualties. It is not a major problem for fatal accidents, because very few fatal accidents do not become known to the police. However, 'Road Casualties Great Britain 2006' (Department for Transport, 2007*c*) suggests that "there is evidence that an appreciable proportion of non-fatal injury accidents are not reported to the police" and "the police tend to underestimate the severity of injury because of the difficulty in distinguishing severity at the scene of the accident". These could cause problems for models of serious and slight accidents because the accident data used for the response variable could have measurement error. In addition to these problems, accidents involving some types of road user like pedal cyclist are particularly likely to be under-reported (see Department for Transport, 2007*c*, Section 6), estimates from the models could be biased. Moreover, if there were a systematic change in the levels of reporting and misclassification, this would cause a more serious problem in monitoring trends in numbers of accidents and casualties. The above problems with STATS19 are common with most accident research in the UK. However, STATS19 remains the best and most complete source of national accident data for research. One approach for assessing the level of under-reporting in STATS19, used by the Department for Transport, is to compare STATS19 data with other sources of data, for instance, the Hospital Episodes Statistics database held by the NHS (Department for Transport, 2006*b*). Further research is needed to assess the effect of under-reporting on road accident models.

To obtain data of good quality and for all desired explanatory variables is difficult. All the data used in this research are secondary data most of which were obtained from

the Department for Transport, the Office for National Statistics, and the Ordnance Survey. The availability of data has a large impact on which explanatory variables were chosen to be included in the models. The response variable used for areal models during 2001 and 2005 is the number of accidents disaggregated by severity at the local authority level for all road classes. However, traffic levels are different for different road classes and the extent of the association between accident frequencies and traffic levels for different road classes may be different. Therefore, ideally it will be more appropriate to develop models for numbers of accidents disaggregated by road class. This will need data of traffic volume for different road classes in each local authority to be known. For models at the ward level, no traffic variable was included in the models because traffic data were not available at the ward level. In order to take account of the variation in traffic levels in the wards, some proxy variables (populations by mode of traveling to work) were used. For junction models in Coventry, the only known explanatory variable is the junction type (a roundabout, a crossing or a T-junction). The traffic flow data at the junctions were not available. Therefore the analysis based on the Coventry dataset is relatively simple and the spatial models for junction accidents have not been extensively studied.

Considering methods used in this research, there are three main limitations. The approach to taking account of spatially structured random effects in this research is to adopt the CAR model. Comparisons of it and conventional accident models were made. However, there are also other possible modelling approaches to account for spatial effects, for instance, the moving average model and the simultaneous autoregressive model. Although their limitations were discussed, how they actually perform for accident data compared with the CAR model was not studied.

The second limitation is that edge effects were not considered for either areal models or network models. For areal models at the local authority level, this should not be a major problem because the study region is the whole of England. However for models of accidents in the West Midlands at the ward level, this can cause some problems. Wards close to the boundaries of the West Midlands may also be spatially dependent on areas outside the boundaries. This can influence the estimates of accident means especially for

the wards close to the boundaries. For models of accidents on a road network, there is also a need to consider the edge effects. For instance, the M1 intersects with other motorways and A-roads. Therefore, neighbouring links for a link on the M1 should include not only its adjacent links on the M1 but also links on another road that intersects with this link. This will make the models more complicated. For areal models, constructing an external buffer zone is one of the most popular approaches for accommodating edge effects. If observations within the buffer zone are available, it is straightforward to fit models with the whole data covering both the study region and the buffer zone. For network models, more complicated methods might be needed to take account of edge effects.

The last limitation is the method used for examining spatial correlation in residuals. Moran's I statistic is appropriate for examining the spatial correlation in a spatial dataset and can be generalized to examine spatial correlation in the residuals from a linear model. However, there has been little work in the past to study how well it works for residuals from a generalized linear model. In the absence of a more appropriate procedure, Moran's I statistic was used for Pearson residuals instead of raw residuals in this research. Value of Moran's I was calculated for Pearson residuals that were based on a point estimate (the posterior mean) of the expected number of accidents (λ) for all the models. Results show that in most cases it was statistically significant (evidence of the existence of spatial correlation) for non-CAR models and nonsignificant for CAR models. However, using the M1 data as an example, the posterior distribution of values of Moran's I suggests that, for a Poisson regression model with log-normal random effects, values of Moran's I in Bayesian residuals were not as big as that in residuals calculated using the posterior mean of λ and were nonsignificant. This is different from what was obtained for one of the areal models, in which Moran's I test leaded to consistent conclusions by using both types of residuals. These indicate that in some occasions results of significance tests for spatial correlation could be very different by using two types of residuals. In a Bayesian context, the examination of Bayesian residuals seems more appropriate. Further research will be needed to understand why results are so different by using these two types of residuals.

One main application of the CAR models is to predict numbers of accidents in the

future. This is straightforward only if spatial effects are constant over time. However, models at the local authority level suggest that the best performing models assume spatial effects vary over time. In the example, in order to make predictions based on previous data, spatial effects were assumed to be constant in two periods of time. The spatial effects for the predicted year were assumed to be same as those in the latter period. But the most appropriate approach would be to investigate how the changes of spatial effects over time can be modelled. Moreover, in the areal models, an additional dummy variable was included to account for the effect of unitary authority. Results show that unitary authorities had more fatal and serious accidents than other local authorities. However, some unitary authorities are urban areas while some are quite rural areas. The approach to treating them as the same in the models by including the dummy variable seems not very appropriate. Reasons for the positive effect of unitary authority on accident frequencies need further investigation.

9.4 Main contributions of the research

The spatial aspects of road accidents and road networks were not extensively considered in the past in road safety studies using a statistical modelling approach. This research is expected to achieve a further step in modelling of the spatial distribution of road accidents.

According to the author's knowledge, this research is the first study trying to use the spatial layout of road networks to develop CAR models for road accidents. Since traffic moves on roads and road accidents happen on the roads, the spatial layout of road networks provides additional information to aid modelling of the spatial distribution of road accidents. This information deserves further consideration to identify and measure the extent of spatial dependence in accident frequencies.

Models developed in this research take account of both spatial effects and temporal effects and therefore fill a gap in the literature. Better estimates of mean accident frequencies are expected to be achieved from these models. This will benefit policy makers and local authorities in the following ways. First, more reliable conclusions about the reduction of accident frequencies over time can be made. Secondly, better predictions of

accident frequencies in the future can be obtained therefore more useful safety policies can be made. In particular, models with spatial effects have other practical uses. For instance, the pattern of spatial distribution of the spatial effects, shown in a map or road network may help to identify the unobserved factors that are associated with high accident frequencies. These models also provide an alternative approach for ranking sites with high risk. Moreover, high-risk sites identified using spatial effects will help safety engineers to find further insights of road network design and urban planning on the occurrence of road accidents and to decide appropriate engineering treatments on selected sites. According to the author's knowledge, using the CAR model for accident prediction and site ranking were proposed here for the first time in road safety research.

9.5 Suggestions for further research

Some problems where further research is needed were highlighted in Section 9.3. There are a number of further directions in which this research might be developed. The method to take account of spatial dependency in accident models has been developed and examined. It was used in areal models and network models. Areal models studied accidents at the local authority level in England and at the ward level in the West Midlands. The next step of the research for areal models could be to use them for accidents at the district level and the ward level in England. The main problem in achieving these will be the difficulty in obtaining traffic data. For network models, only one motorway and major junctions in an urban area were studied. In the future, the CAR model could be used to study accidents on roads of other classes, for instance A- and B- roads, and at junctions on road networks in other locations. The effects of taking account of spatial dependency in the models for these classes of roads and junctions could be studied. Again, the main problem for achieving these is the difficulty of getting traffic data. For instance, traffic data are usually not available for individual links on B- roads and minor roads. Moreover, a road network is composed of junctions and links. Junction models and link models were developed separately in this research. The approach to jointly modelling junction and link accidents will be investigated. Not only the neighbouring junctions can be spatially cor-

related and the neighbouring links can be spatially correlated, but also a junction can be spatially correlated with its neighbouring links and a link can be spatially correlated with its neighbouring junctions. Therefore by borrowing information from both neighbouring junctions and links, a better estimate of the accident mean at a site could be achieved.

When using a statistical modelling approach, accident data need to be aggregated over space first. The spatial unit for aggregation is usually a local authority, ward, link or junction. There is another type of spatial unit that may be appropriate for analysing accident data. Figure 9.1 has appeared in Chapter 4 to illustrate how a road network can be represented by a node-link graph and how a neighbours list can be identified for such a road network. Areas bounded by the black lines, representing the major roads, are known as *cells*. Methods to model accidents in such cells needs to be considered. How accident frequencies in neighbouring cells can be spatially correlated and how they can be related to accident frequencies on the major roads that bound the cells need to be studied in the future.

Figure 9.1: Accidents on a road network.

How the CAR model can be used to rank high-risk sites based on the estimates of spatial effects was introduced. Suppose that some high-risk sites identified by this ap-

proach had been selected for engineering treatment. By using data for accidents and other explanatory variables in the year after the treatment, spatial effects at all sites could be estimated. How they would vary before and after the treatment at treated sites as well as at untreated sites, especially those close to treated sites, would be an interesting question to be answered. If the treatment is successful, the level of spatial effects is expected to become small at treated sites and not to increase at untreated sites that are close to treated sites. If the level of spatial effects becomes small at treated sites but the level of spatial effects is found to be much larger at their neighbouring untreated sites after the treatment, this may indicate an accident migration problem. Therefore, further research will be made for investigating the practical use of the CAR model to examine the effect of engineering treatment and to provide a better understanding of accident migration. However, there are some difficulties to achieve this in practice. Firstly, an appropriate road network needs to be selected and it should contain several high-risk sites identified by the approach proposed in this research. Secondly, on-site investigation will be needed to confirm contributory factors for high spatial effects at these sites and to decide remedial measures on the selected sites. Thirdly, the treatment needs to be agreed by the local authority who takes the responsibility of the road network. Lastly, the road network needs to be monitored over a period of time and all necessary data need to be recorded both before and after the treatment.

With regard to statistical methods, there are at least two aspects that need to be considered or further developed in the future. Firstly, a better approach or an improved approach based on Moran's I statistic needs to be developed for examining spatial correlation in the residuals from a generalized linear model. Secondly, methods for posterior predictive checks of Bayesian models need to be further developed, especially those for examining spatial models. Previous applications of Bayesian models, including those in spatial epidemiology, seldom seem to have done such model checking. However, it is an important procedure to examine the validity of the model.

Spatial statistics and modelling approaches have a long developing history and have been well implemented in some research areas. Although models with spatial effects

can be complicated, the computation is relatively easy by using the Bayesian approach. Recent development of geographic information systems (GIS) provides a powerful tool to integrate different types of spatial data. These give better opportunities to study the spatial distribution of road accidents.

Appendix A

Lists of local authorities and wards

Table A.1: Lists of local authorities in Englands in the 1980s

Index	Local authority	Metropolitan county
1	Lincolnshire	
2	Cumbria	
3	North Yorkshire	
4	Northumberland	
5	Cornwall	
6	Devon	
7	Somerset	
8	Dorset	
9	East Sussex	
10	Wiltshire	
11	Hampshire	
12	Berkshire	
13	Hereford and Worcester	
14	Gloucestershire	
15	Oxfordshire	
16	Buckinghamshire	
17	Warwickshire	
18	Kent	
19	Hertfordshire	
20	Northamptonshir	
21	Cambridgeshire	
22	Essex	
23	Suffolk	
24	Shropshire	
25	Leicestershire	

Continued on next page...

Table A.1 – Continued

Index	Local authority	Metropolitan county
26	Staffordshire	
27	Derbyshire	
28	Nottinghamshire	
29	Cheshire	
30	Lancashire	
31	Norfolk	
32	Durham	
33	Isle of Wight	
34	West Sussex	
35	Avon	
36	Surrey	
37	Bedfordshire	
38	Cleveland	
39	Wigan	Greater Manchester
40	Kirklees	West Yorkshire
41	Calderdale	West Yorkshire
42	Bradford	West Yorkshire
43	Doncaster	South Yorkshire
44	Leeds	West Yorkshire
45	Wakefield	West Yorkshire
46	Gateshead	Tyne and Wear
47	Liverpool	Merseyside
48	Sefton	Merseyside
49	Dudley	West Midlands
50	Solihull	West Midlands
51	Birmingham	West Midlands
52	Walsall	West Midlands
53	Coventry	West Midlands
54	Bromley	Greater London
55	Richmond upon Thames	Greater London
56	Hillingdon	Greater London
57	Havering	Greater London
58	Knowsley	Merseyside
59	St Helens	Merseyside
60	Trafford	Greater Manchester
61	Manchester	Greater Manchester
62	Salford	Greater Manchester
63	Tameside	Greater Manchester
64	Sheffield	South Yorkshire
65	Rotherham	South Yorkshire

Continued on next page...

Table A.1 – Continued

Index	Local authority	Metropolitan county
66	Bolton	Greater Manchester
67	Bury	Greater Manchester
68	Oldham	Greater Manchester
69	Rochdale	Greater Manchester
70	Barnsley	South Yorkshire
71	Sunderland	Tyne and Wear
72	South Tyneside	Tyne and Wear
73	Wirral	Merseyside
74	Wolverhampton	West Midlands
75	Sandwell	West Midlands
76	Kingston upon Thames	Greater London
77	Sutton	Greater London
78	Hounslow	Greater London
79	Merton	Greater London
80	Wandsworth	Greater London
81	Croydon	Greater London
82	Lambeth	Greater London
83	Southwark	Greater London
84	Lewisham	Greater London
85	Greenwich	Greater London
86	Ealing	Greater London
87	Hammersmith and Fulham	Greater London
88	Brent	Greater London
89	Harrow	Greater London
90	Barnet	Greater London
91	Islington	Greater London
92	Hackney	Greater London
93	Newham	Greater London
94	Barking and Dagen	Greater London
95	Haringey	Greater London
96	Enfield	Greater London
97	Waltham Forest	Greater London
98	Redbridge	Greater London
99	Bexley	Greater London
100	Stockport	Greater Manchester
101	Newcastle-upon-Tyne	Tyne and Wear
102	North Tyneside	Tyne and Wear
103	Kensington and Chelsea	Greater London
104	Westminster, City of	Greater London
105	Camden	Greater London

Continued on next page...

Table A.1 – Continued

Index	Local authority	Metropolitan county
106	Tower Hamlets	Greater London
107	City of London	Greater London
108	Humberside	

Table A.2: Lists of local authorities in Englands in the 2000s

Index	Local authority	Metropolitan county
1	Lincolnshire	
2	Cumbria	
3	North Yorkshire	
4	Northumberland	
5	Cornwall	
6	Devon	
7	Somerset	
8	Dorset	
9	East Sussex	
10	Wiltshire	
11	Hampshire	
12	Gloucestershire	
13	Oxfordshire	
14	Warwickshire	
15	Dudley	West Midlands
16	Solihull	West Midlands
17	Birmingham	West Midlands
18	Walsall	West Midlands
19	Coventry	West Midlands
20	Wolverhampton	West Midlands
21	Sandwell	West Midlands
22	Bromley	Greater London
23	Richmond upon Thames	Greater London
24	Hillingdon	Greater London
25	Havering	Greater London
26	Kingston upon Thames	Greater London
27	Sutton	Greater London
28	Hounslow	Greater London
29	Merton	Greater London
30	Wandsworth	Greater London
31	Croydon	Greater London
32	Lambeth	Greater London
33	Southwark	Greater London
34	Lewisham	Greater London
35	Greenwich	Greater London
36	Ealing	Greater London
37	Hammersmith and Fulham	Greater London
38	Brent	Greater London
39	Harrow	Greater London
40	Barnet	Greater London

Continued on next page...

Table A.2 – Continued

Index	Local authority	Metropolitan county
41	Islington	Greater London
42	Hackney	Greater London
43	Newham	Greater London
44	Barking and Dagenham	Greater London
45	Haringey	Greater London
46	Enfield	Greater London
47	Waltham Forest	Greater London
48	Redbridge	Greater London
49	Bexley	Greater London
50	Kensington and Chelsea	Greater London
51	Westminster	Greater London
52	Camden	Greater London
53	Tower Hamlets	Greater London
54	City of London	Greater London
55	Kent	
56	Hertfordshire	
57	Northamptonshire	
58	Cambridgeshire	
59	Essex	
60	Suffolk	
61	Shropshire	
62	Leicestershire	
63	Staffordshire	
64	Derbyshire	
65	Nottinghamshire	
66	Cheshire	
67	Lancashire	
68	Doncaster	South Yorkshire
69	Sheffield	South Yorkshire
70	Rotherham	South Yorkshire
71	Barnsley	South Yorkshire
72	Kirklees	West Yorkshire
73	Calderdale	West Yorkshire
74	Bradford	West Yorkshire
75	Leeds	West Yorkshire
76	Wakefield	West Yorkshire
77	Norfolk	
78	North Lincolnshire (UA)	
79	East Riding of Yorkshire (UA)	
80	Durham	

Continued on next page...

Table A.2 – Continued

Index	Local authority	Metropolitan county
81	Isle of Wight (UA)	
82	West Sussex	
83	North Somerset (UA)	
84	Bristol and City of (UA)	
85	West Berkshire (UA)	
86	Wokingham (UA)	
87	Buckinghamshire	
88	Herefordshire (UA)	
89	Worcestershire	
90	Surrey	
91	Bedfordshire	
92	Peterborough (UA)	
93	Telford and Wrekin (UA)	
94	Wirral	Merseyside
95	Liverpool	Merseyside
96	Sefton	Merseyside
97	Knowsley	Merseyside
98	St. Helens	Merseyside
99	Wigan	Greater Manchester
100	Trafford	Greater Manchester
101	Manchester	Greater Manchester
102	Salford	Greater Manchester
103	Tameside	Greater Manchester
104	Bolton	Greater Manchester
105	Bury	Greater Manchester
106	Rochdale	Greater Manchester
107	Oldham	Greater Manchester
108	Stockport	Greater Manchester
109	Gateshead	Tyne and Wear
110	Sunderland	Tyne and Wear
111	South Tyneside	Tyne and Wear
112	Newcastle upon Tyne	Tyne and Wear
113	North Tyneside	Tyne and Wear
114	Plymouth (UA)	
115	Torbay (UA)	
116	Blackpool (UA)	
117	Poole (UA)	
118	Bournemouth (UA)	
119	Southampton (UA)	
120	Portsmouth (UA)	

Continued on next page...

Table A.2 – Continued

Index	Local authority	Metropolitan county
121	Brighton and Hove (UA)	
122	Bath and North East Somerset (UA)	
123	South Gloucestershire (UA)	
124	Swindon (UA)	
125	Bracknell Forest (UA)	
126	Reading (UA)	
127	Windsor and Maidenhead (UA)	
128	Slough (UA)	
129	Milton Keynes (UA)	
130	Leicester (UA)	
131	Rutland (UA)	
132	Medway (UA)	
133	Thurrock (UA)	
134	Luton (UA)	
135	Southend-on-Sea (UA)	
136	Stoke-on-Trent (UA)	
137	Halton (UA)	
138	Warrington (UA)	
139	Derby (UA)	
140	Nottingham (UA)	
141	Blackburn with Darwen (UA)	
142	York (UA)	
143	North East Lincolnshire (UA)	
144	Kingston upon Hull and City of (UA)	
145	Darlington (UA)	
146	Stockton-on-Tees (UA)	
147	Middlesbrough (UA)	
148	Hartlepool (UA)	
149	Redcar and Cleveland (UA)	

Table A.3: Lists of wards in the West Midlands

Index	Ward	District
1	Acock's Green	Birmingham
2	Aston	Birmingham
3	Bartley Green	Birmingham
4	Billesley	Birmingham
5	Bournville	Birmingham
6	Brandwood	Birmingham
7	Edgbaston	Birmingham
8	Erdington	Birmingham
9	Fox Hollies	Birmingham
10	Hall Green	Birmingham
11	Handsworth	Birmingham
12	Harborne	Birmingham
13	Hodge Hill	Birmingham
14	Kingsbury	Birmingham
15	King's Norton	Birmingham
16	Kingstanding	Birmingham
17	Ladywood	Birmingham
18	Longbridge	Birmingham
19	Moseley	Birmingham
20	Nechells	Birmingham
21	Northfield	Birmingham
22	Oscott	Birmingham
23	Perry Barr	Birmingham
24	Quinton	Birmingham
25	Sandwell	Birmingham
26	Selly Oak	Birmingham
27	Shard End	Birmingham
28	Sheldon	Birmingham
29	Small Heath	Birmingham
30	Soho	Birmingham
31	Sparkbrook	Birmingham
32	Sparkhill	Birmingham
33	Stockland Green	Birmingham
34	Sutton Four Oaks	Birmingham
35	Sutton New Hall	Birmingham
36	Sutton Vesey	Birmingham
37	Washwood Heath	Birmingham
38	Weoley	Birmingham
39	Yardley	Birmingham
40	Bablake	Coventry

Continued on next page...

Table A.3 – Continued

Index	Ward	District
41	Binley and Willenhall	Coventry
42	Cheylesmore	Coventry
43	Earlsdon	Coventry
44	Foleshill	Coventry
45	Henley	Coventry
46	Holbrook	Coventry
47	Longford	Coventry
48	Lower Stoke	Coventry
49	Radford	Coventry
50	St. Michael's	Coventry
51	Sherbourne	Coventry
52	Upper Stoke	Coventry
53	Wainbody	Coventry
54	Westwood	Coventry
55	Whoberley	Coventry
56	Woodlands	Coventry
57	Wyken	Coventry
58	Amblecote	Dudley
59	Belle Vale and Hasbury	Dudley
60	Brierley Hill	Dudley
61	Brockmoor and Pensnett	Dudley
62	Castle and Priory	Dudley
63	Coseley East	Dudley
64	Coseley West	Dudley
65	Gornal	Dudley
66	Halesowen North	Dudley
67	Halesowen South	Dudley
68	Hayley Green	Dudley
69	Kingswinford North and Wall Heath	Dudley
70	Kingswinford South	Dudley
71	Lye and Wollescote	Dudley
72	Netherton and Woodside	Dudley
73	Norton	Dudley
74	Pedmore and Stourbridge East	Dudley
75	Quarry Bank and Cradley	Dudley
76	St. Andrews	Dudley
77	St. James's	Dudley
78	St. Thomas's	Dudley
79	Sedgley	Dudley
80	Wollaston and Stourbridge West	Dudley

Continued on next page...

Table A.3 – Continued

Index	Ward	District
81	Wordsley	Dudley
82	Abbey	Sandwell
83	Blackheath	Sandwell
84	Bristnall	Sandwell
85	Charlemont	Sandwell
86	Cradley Heath and Old Hill	Sandwell
87	Friar Park	Sandwell
88	Great Barr	Sandwell
89	Great Bridge	Sandwell
90	Greets Green and Lyng	Sandwell
91	Hateley Heath	Sandwell
92	Langley	Sandwell
93	Newton	Sandwell
94	Oldbury	Sandwell
95	Old Warley	Sandwell
96	Princes End	Sandwell
97	Rowley	Sandwell
98	St. Pauls	Sandwell
99	Smethwick	Sandwell
100	Soho and Victoria	Sandwell
101	Tipton Green	Sandwell
102	Tividale	Sandwell
103	Wednesbury North	Sandwell
104	Wednesbury South	Sandwell
105	West Bromwich Central	Sandwell
106	Bickenhill	Solihull
107	Castle Bromwich	Solihull
108	Chelmsley Wood	Solihull
109	Elmdon	Solihull
110	Fordbridge	Solihull
111	Kingshurst	Solihull
112	Knowle	Solihull
113	Lyndon	Solihull
114	Meriden	Solihull
115	Olton	Solihull
116	Packwood	Solihull
117	St. Alphege	Solihull
118	Shirley East	Solihull
119	Shirley South	Solihull
120	Shirley West	Solihull

Continued on next page...

Table A.3 – Continued

Index	Ward	District
121	Silhill	Solihull
122	Smith's Wood	Solihull
123	Aldridge Central and South	Walsall
124	Aldridge North and Walsall Wood	Walsall
125	Bentley and Darlaston North	Walsall
126	Birchills Leamore	Walsall
127	Blakenall	Walsall
128	Bloxwich East	Walsall
129	Bloxwich West	Walsall
130	Brownhills	Walsall
131	Darlaston South	Walsall
132	Hatherton Rushall	Walsall
133	Paddock	Walsall
134	Palfrey	Walsall
135	Pelsall	Walsall
136	Pheasey	Walsall
137	Pleck	Walsall
138	St. Matthew's	Walsall
139	Short Heath	Walsall
140	Streetly	Walsall
141	Willenhall North	Walsall
142	Willenhall South	Walsall
143	Bilston East	Wolverhampton
144	Bilston North	Wolverhampton
145	Blakenhall	Wolverhampton
146	Bushbury	Wolverhampton
147	East Park	Wolverhampton
148	Ettingshall	Wolverhampton
149	Fallings Park	Wolverhampton
150	Graiseley	Wolverhampton
151	Heath Town	Wolverhampton
152	Low Hill	Wolverhampton
153	Merry Hill	Wolverhampton
154	Oxley	Wolverhampton
155	Park	Wolverhampton
156	Penn	Wolverhampton
157	St. Peter's	Wolverhampton
158	Spring Vale	Wolverhampton
159	Tettenhall Regis	Wolverhampton
160	Tettenhall Wightwick	Wolverhampton

Continued on next page...

Table A.3 – Continued

Index	Ward	District
161	Wednesfield North	Wolverhampton
162	Wednesfield South	Wolverhampton

Appendix B

Parameter estimates for selected areal models (1983 - 1986)

Table B.1: Parameter estimates for fatal accidents

Parameter	2.5%	Median	97.5%	$\hat{R}$
Model PL				
Intercept	2.27	2.30	2.34	1.001
Area	-0.48	-0.35	-0.23	1.003
Population	0.33	0.50	0.67	1.007
Vehicle	-0.43	-0.22	-0.05	1.011
Road	0.05	0.20	0.36	1.021
Traffic	0.23	0.36	0.49	1.031
Model PLfe				
Intercept	2.06	2.18	2.30	1.000
Area	-0.40	-0.28	-0.15	1.006
Population	0.49	0.71	0.92	1.002
Vehicle	-0.66	-0.43	-0.19	1.003
Road	0.12	0.27	0.43	1.004
Traffic	0.18	0.31	0.44	1.009
Metropolitan district	-0.11	0.05	0.21	1.000
London borough	0.12	0.34	0.55	1.000
Model PLre				
Intercept	2.22	2.35	2.49	1.011
Area	-0.51	-0.37	-0.23	1.000
Population	0.70	0.95	1.19	1.002

Continued on next page...

Table B.1 – Continued

Parameter	2.5%	Median	97.5%	$\hat{R}$
Vehicle	-0.86	-0.62	-0.36	1.006
Road	0.04	0.19	0.36	1.000
Traffic	0.17	0.29	0.43	1.004
London borough	-0.18	0.06	0.31	1.009
Greater Manchester	-0.44	-0.23	-0.01	1.016
Tyne and Wear	-1.10	-0.75	-0.42	1.006
West Yorkshire	-0.10	0.07	0.24	1.013
South Yorkshire	-0.33	-0.12	0.09	1.002
Merseyside	-0.52	-0.25	-0.01	1.005
West Midlands	-0.33	-0.13	0.07	1.004
Model PLNre				
Intercept	2.21	2.37	2.53	1.019
Area	-0.53	-0.35	-0.19	1.002
Population	0.63	0.91	1.23	1.022
Vehicle	-0.97	-0.62	-0.31	1.016
Road	-0.02	0.17	0.35	1.016
Traffic	0.15	0.32	0.47	1.002
London borough	-0.3	0.01	0.29	1.005
Greater Manchester	-0.54	-0.28	-0.03	1.037
Tyne and Wear	-1.18	-0.79	-0.43	1.060
West Yorkshire	-0.14	0.06	0.28	1.033
South Yorkshire	-0.40	-0.11	0.14	1.038
Merseyside	-0.61	-0.30	0.02	1.090
West Midlands	-0.44	-0.20	0.03	1.009
σ_v[1]	0.15	0.20	0.25	1.008

[1] σ_v: standard deviation of unstructured random effects.

Table B.2: Parameter estimates for serious accidents

Parameter	2.5%	Median	97.5%	$\hat{R}$
Model PL				
Intercept	4.84	4.85	4.86	1.005
Area	-0.35	-0.32	-0.29	1.011
Population	-0.39	-0.33	-0.29	1.033
Vehicle	0.72	0.77	0.82	1.027
Road	0.09	0.14	0.18	1.015
Traffic	0.22	0.25	0.29	1.012
Model PLfe				
Intercept	4.73	4.77	4.80	1.008
Area	-0.31	-0.28	-0.24	1.002
Population	-0.03	0.04	0.09	1.011
Vehicle	0.33	0.39	0.45	1.013
Road	0.23	0.27	0.32	1.000
Traffic	0.13	0.17	0.21	1.000
Metropolitan district	-0.15	-0.11	-0.06	1.013
London borough	0.27	0.33	0.39	1.001
Model PLre				
Intercept	4.80	4.84	4.88	1.000
Area	-0.31	-0.26	-0.23	1.000
Population	0.13	0.19	0.26	1.019
Vehicle	0.09	0.17	0.25	1.007
Road	0.21	0.25	0.30	1.000
Traffic	0.16	0.20	0.23	1.000
London borough	0.18	0.24	0.32	1.000
Greater Manchester	-0.39	-0.32	-0.26	1.000
Tyne and Wear	-0.43	-0.34	-0.25	1.000
West Yorkshire	-0.17	-0.12	-0.06	1.007
South Yorkshire	-0.37	-0.29	-0.23	1.002
Merseyside	-0.46	-0.38	-0.30	1.000
West Midlands	-0.10	-0.04	0.01	1.000
Model PLN				
Intercept	4.77	4.80	4.82	1.000
Area	-0.39	-0.27	-0.16	1.098
Population	-0.82	-0.59	-0.43	1.093
Vehicle	0.71	0.89	1.11	1.127
Road	0.05	0.18	0.31	1.115
Traffic	0.19	0.31	0.44	1.098
σ_v	0.28	0.31	0.33	1.004

Continued on next page...

Table B.2 – Continued

Parameter	2.5%	Median	97.5%	$\hat{R}$
Model $ICAR_{nb1}$				
Intercept	4.78	4.79	4.80	1.001
Area	-0.28	-0.18	-0.08	1.029
Population	-0.12	0.07	0.25	1.026
Vehicle	-0.01	0.16	0.38	1.021
Road	0.11	0.24	0.36	1.05
Traffic	0.18	0.28	0.39	1.08
$\sigma_{\theta 1}$[1]	0.44	0.52	0.62	1.001
$\sigma_{\theta 2}$	0.46	0.55	0.64	1.000
$\sigma_{\theta 3}$	0.45	0.52	0.60	1.005
$\sigma_{\theta 4}$	0.45	0.52	0.62	1.000
Model $CCAR_{nb1}$				
Intercept	4.78	4.79	4.81	1.000
Area	-0.24	-0.16	-0.05	1.000
Population	-0.31	-0.07	2.32	1.216
Vehicle	-0.02	0.29	0.54	1.192
Road	0.10	0.22	0.35	1.012
Traffic	0.22	0.31	0.41	1.022
σ_v	0.12	0.16	0.19	1.001
$\sigma_{\theta 1}$	0.29	0.40	0.52	1.004
$\sigma_{\theta 2}$	0.31	0.43	0.54	1.000
$\sigma_{\theta 3}$	0.32	0.41	0.52	1.001
$\sigma_{\theta 4}$	0.34	0.43	0.54	1.000
Model $CCAR_{nb3road}$				
Intercept	4.78	4.79	4.81	1.005
Area	-0.26	-0.17	-0.06	1.019
Population	-0.19	.03	0.23	1.003
Vehicle	0.07	0.29	0.58	1.009
Road	0.10	0.21	0.33	1.010
Traffic	0.15	0.25	0.37	1.010
σ_v	0.14	0.17	0.20	1.005
$\sigma_{\theta 1}$	0.33	0.44	0.57	1.000
$\sigma_{\theta 2}$	0.35	0.47	0.62	1.002
$\sigma_{\theta 3}$	0.36	0.46	0.57	1.001
$\sigma_{\theta 4}$	0.38	0.48	0.62	1.007

[1] σ_θ: standard deviation of spatially structured random effects

Table B.3: Parameter estimates for slight accidents

Parameter	2.5%	Median	97.5%	$\hat{R}$
Model PL				
Intercept	6.17	6.18	6.19	1.000
Area	-0.38	-0.37	-0.35	1.017
Population	-0.55	-0.52	-0.50	1.004
Vehicle	0.61	0.64	0.67	1.007
Road	0.01	0.03	0.05	1.007
Traffic	0.55	0.57	0.58	1.015
Model PLfe				
Intercept	6.39	6.41	6.43	1.005
Area	-0.50	-0.49	-0.47	1.005
Population	0.26	0.30	0.32	1.019
Vehicle	-0.30	-0.27	-0.23	1.025
Road	0.18	0.21	0.23	1.000
Traffic	0.48	0.49	0.51	1.000
Metropolitan district	-0.67	-0.64	-0.62	1.004
London borough	-0.13	-0.10	-0.07	1.007
Model PLre				
Intercept	6.39	6.41	6.42	1.038
Area	-0.55	-0.53	-0.51	1.011
Population	0.21	0.24	0.28	1.020
Vehicle	-0.16	-0.12	-0.08	1.028
Road	0.15	0.17	0.19	1.003
Traffic	0.44	0.46	0.48	1.004
London borough	-0.15	-0.12	-0.08	1.035
Greater Manchester	-0.53	-0.51	-0.47	1.038
Tyne and Wear	-0.89	-0.85	-0.80	1.044
West Yorkshire	-0.60	-0.57	-0.54	1.013
South Yorkshire	-0.58	-0.54	-0.51	1.028
Merseyside	-0.54	-0.51	-0.46	1.023
West Midlands	-0.87	-0.83	-0.80	1.017
Model PLN				
Intercept	6.11	6.14	6.17	1.034
Area	-0.26	-0.16	0.15	1.041
Population	-1.05	-0.70	-0.16	1.282
Vehicle	-0.08	0.56	0.77	1.172
Road	-0.49	-0.09	0.12	1.01
Traffic	0.65	0.76	1.15	1.074
σ_v	0.31	0.33	0.38	1.016

Continued on next page...

Table B.3 – Continued

Parameter	2.5%	Median	97.5%	$\hat{R}$
Model $CCAR_{nb3road}$				
Intercept	6.12	6.14	6.15	1.002
Area	-0.26	-0.19	-0.07	1.004
Population	-0.39	0.11	0.27	1.256
Vehicle	-0.06	0.08	0.39	1.211
Road	-0.12	0.07	0.16	1.029
Traffic	0.41	0.48	0.66	1.070
σ_v	0.10	0.13	0.16	1.11
$\sigma_{\theta 1}$	0.44	0.53	0.64	1.024
$\sigma_{\theta 2}$	0.41	0.51	0.63	1.004
$\sigma_{\theta 3}$	0.43	0.54	0.66	1.008
$\sigma_{\theta 4}$	0.40	0.50	0.61	1.006

Appendix C

Parameter estimates for selected areal models (2001 - 2005)

Table C.1: Model PLtr

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs[1])	5.08	5.09	5.10	1.002
Intercept (sl[2])	6.87	6.87	6.87	1.000
Trend (fs)	-0.07	-0.07	-0.07	1.001
Trend (sl)	-0.05	-0.04	-0.04	1.002
Area (fs)	-0.16	-0.13	-0.10	1.000
Area (sl)	-0.24	-0.23	-0.22	1.002
Population (fs)	0.68	0.71	0.73	1.005
Population (sl)	0.51	0.52	0.53	1.005
A-road (fs)	0.45	0.48	0.51	1.008
A-road (sl)	0.23	0.24	0.25	1.004
B-road (fs)	0.06	0.08	0.09	1.000
B-road (sl)	0.04	0.05	0.06	1.000
Minor road (fs)	-0.47	-0.44	-0.41	1.003
Minor road (sl)	-0.03	-0.02	-0.01	1.003
Other traffic (fs)	0.36	0.40	0.43	1.001
Other traffic (sl)	0.22	0.23	0.25	1.003
Car traffic (fs)	-0.39	-0.36	-0.32	1.001
Car traffic (sl)	-0.08	-0.07	-0.06	1.003
Junction (fs)	-0.02	-0.01	0.01	1.006
Junction (sl)	0.15	0.16	0.17	1.000

[1] 'fs': fatal and serious accidents.
[2] 'sl': for slight accidents.

Table C.2: Model PLtr-fe

Parameter	2.5%	Median	97.5%	$\hat{R}$
PLtr-fe				
Intercept (fs)	4.90	4.92	4.05	1.000
Intercept (sl)	6.79	6.80	6.81	1.003
Trend (fs)	-0.07	-0.07	-0.06	1.001
Trend (sl)	-0.05	-0.04	-0.04	1.000
London borough (fs)	0.40	0.44	0.48	1.002
London borough (sl)	0.16	0.18	0.20	1.003
Metropolitan district (fs)	-0.07	-0.04	-0.01	1.005
Metropolitan district (sl)	-0.02	-0.01	-0.001	1.002
Unitary authority (fs)	0.19	0.22	0.26	1.002
Unitary authority (sl)	0.05	0.06	0.08	1.001
Area (fs)	-0.17	-0.14	-0.11	1.003
Area (sl)	-0.25	-0.24	-0.23	1.002
Population (fs)	0.45	0.48	0.51	1.012
Population (sl)	0.41	0.42	0.43	1.000
A-road (fs)	0.35	0.38	0.41	1.011
A-road (sl)	0.18	0.19	0.20	1.00
B-road (fs)	0.12	0.14	0.17	1.003
B-road (sl)	0.07	0.08	0.09	1.002
Minor road (fs)	-0.13	-0.09	-0.05	1.024
Minor road (sl)	0.12	0.14	0.15	1.000
Other traffic (fs)	0.36	0.40	0.44	1.003
Other traffic (sl)	0.21	0.23	0.24	1.008
Car traffic (fs)	-0.27	-0.24	-0.20	1.003
Car traffic (sl)	-0.03	-0.01	0.002	1.012
Junction (fs)	0.13	0.15	0.16	1.003
Junction (sl)	0.21	0.22	0.22	1.002

Table C.3: Model PLtr-re

Parameter	2.5%	Median	97.5%	$\hat{R}$
Trend (fs)	-0.07	-0.07	-0.06	1.001
Trend (sl)	-0.04	-0.04	-0.04	1.002
Intercept (fs)	4.93	4.96	4.99	1.002
Intercept (sl)	6.79	6.81	6.82	1.001
London borough (fs)	0.37	0.41	0.46	1.002
London borough (sl)	0.18	0.19	0.21	1.000
West Midlands (fs)	-0.04	-0.00	0.04	1.000
West Midlands (sl)	0.04	0.06	0.08	1.001
Greater Manchester (fs)	-0.31	-0.27	-0.23	1.000
Greater Manchester (sl)	-0.04	-0.03	-0.01	1.003
West Yorkshire (fs)	-0.08	-0.04	-0.01	1.000
West Yorkshire (sl)	-0.07	-0.05	-0.04	1.002
South Yorkshire (fs)	-0.12	-0.07	-0.03	1.000
South Yorkshire (sl)	-0.04	-0.02	-0.01	1.004
Merseyside (fs)	0.01	0.05	0.10	1.001
Merseyside (sl)	-0.03	-0.01	0.00	1.002
Tyne and Wear (fs)	-0.33	-0.28	-0.22	1.005
Tyne and Wear (sl)	-0.20	-0.17	-0.14	1.000
Area (fs)	-0.18	-0.14	-0.11	1.008
Area (sl)	-0.25	-0.24	-0.23	1.007
Population (fs)	0.37	0.40	0.43	1.008
Population (sl)	0.36	0.37	0.39	1.003
A-road (fs)	0.34	0.37	0.40	1.044
A-road (sl)	0.19	0.20	0.22	1.002
B-road (fs)	0.14	0.16	0.18	1.000
B-road (sl)	0.07	0.08	0.09	1.002
Minor road (fs)	-0.09	-0.04	-0.01	1.010
Minor road (sl)	0.16	0.17	0.19	1.001
Other traffic (fs)	0.34	0.38	0.41	1.001
Other traffic (sl)	0.19	0.20	0.22	1.010
Car traffic (fs)	-0.22	-0.18	-0.14	1.002
Car traffic (sl)	0.01	0.02	0.04	1.005
Junction (fs)	0.17	0.19	0.21	1.000
Junction (sl)	0.24	0.25	0.26	1.007

Table C.4: Model PLNtr

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs)	5.08	5.11	5.14	1.000
Intercept (sl)	6.86	6.88	6.91	1.000
Trend (fs)	-0.09	-0.08	-0.07	1.000
Trend (sl)	-0.06	-0.05	-0.04	1.000
Area (fs)	-0.25	-0.17	-0.09	1.039
Area (sl)	-0.38	-0.30	-0.23	1.042
Population (fs)	0.48	0.52	0.57	1.040
Population (sl)	0.31	0.35	0.39	1.048
A-road (fs)	0.51	0.57	0.65	1.013
A-road (sl)	0.28	0.33	0.41	1.032
B-road (fs)	-0.04	0.02	0.08	1.059
B-road (sl)	-0.04	0.01	0.06	1.088
Minor road (fs)	-0.46	-0.36	-0.28	1.085
Minor road (sl)	-0.02	0.07	0.14	1.081
Other traffic (fs)	0.51	0.59	0.66	1.062
Other traffic (sl)	0.36	0.42	0.48	1.077
Car traffic (fs)	-0.50	-0.43	-0.35	1.006
Car traffic (sl)	-0.22	-0.16	-0.09	1.015
Junction (fs)	0.02	0.06	0.10	1.089
Junction (sl)	0.17	0.21	0.25	1.081
σ_v^{21}	0.03	0.04	0.04	1.003

[1] σ_v^2: variance of unstructured random effects.

Table C.5: Model CCAR(t)tr.temp

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs)	5.00	5.10	5.23	1.025
Intercept (sl)	6.73	6.87	7.03	1.089
Trend (fs)	-0.10	-0.09	-0.07	1.015
Trend (sl)	-0.07	-0.05	-0.03	1.092
Unitary authority (fs)	0.06	0.11	0.18	1.006
Unitary authority (sl)	-0.01	0.03	0.07	1.000
Area (fs)	-0.35	-0.24	-0.14	1.019
Area (sl)	-0.33	-0.26	-0.20	1.060
Population (fs)	0.16	0.22	0.28	1.025
Population (sl)	0.10	0.16	0.20	1.919
A-road (fs)	0.40	0.49	0.57	1.139
A-road (sl)	0.14	0.20	0.27	1.378
B-road (fs)	0.02	0.08	0.13	1.004
B-road (sl)	0.02	0.06	0.10	1.098
Minor road (fs)	0.13	0.25	0.38	1.071
Minor road (sl)	0.36	0.46	0.59	1.717
Other traffic (fs)	0.24	0.36	0.48	1.027
Other traffic (sl)	0.25	0.33	0.42	1.570
Car traffic (fs)	-0.36	-0.23	-0.12	1.028
Car traffic (sl)	-0.17	-0.07	-0.01	1.458
Junction (fs)	0.19	0.25	0.32	1.017
Junction (sl)	0.34	0.38	0.43	1.735
σ_v^2 (fs)	0.019	0.026	0.032	1.008
σ_v^2 (sl)	0.014	0.017	0.021	1.237
$\sigma_\theta^{2\,1}$ (fs, 2001)	0.065	0.106	0.159	1.000
σ_θ^2 (fs, 2002)	0.055	0.094	0.146	1.003
σ_θ^2 (fs, 2003)	0.039	0.070	0.115	1.002
σ_θ^2 (fs, 2004)	0.032	0.063	0.101	1.012
σ_θ^2 (fs, 2005)	0.029	0.061	0.104	1.022
σ_θ^2 (sl, 2001)	0.021	0.038	0.062	1.066
σ_θ^2 (sl, 2002)	0.020	0.041	0.067	1.075
σ_θ^2 (sl, 2003)	0.015	0.032	0.058	1.118
σ_θ^2 (sl, 2004)	0.010	0.024	0.044	1.027
σ_θ^2 (sl, 2005)	0.011	0.025	0.044	1.073

[1] σ_θ^2: overall variance of spatially structured random effects.

Table C.6: Model MVCCAR(t)tr.temp.mv

Parameter	2.5%	Median	97.5%	$\hat{R}$
$\sigma^2_{v.fs}$	0.013	0.022	0.031	1.008
$\sigma^2_{v.sl}$	0.008	0.012	0.016	1.006
$\sigma_{v.fs}\sigma_{v.sl}$	0.003	0.007	0.011	1.002
$\sigma^2_{\theta.fs}$ (2001)	0.072	0.112	0.163	1.000
$\sigma^2_{\theta.sl}$ (2001)	0.027	0.044	0.067	1.000
$\sigma_{\theta.fs}\sigma_{\theta.sl}$ (2001)	0.032	0.055	0.084	1.000
$\sigma^2_{\theta.fs}$ (2002)	0.060	0.102	0.154	1.007
$\sigma^2_{\theta.sl}$ (2002)	0.027	0.047	0.072	1.006
$\sigma_{\theta.fs}\sigma_{\theta.sl}$ (2002)	0.029	0.054	0.082	1.010
$\sigma^2_{\theta.fs}$ (2003)	0.042	0.075	0.118	1.002
$\sigma^2_{\theta.sl}$ (2003)	0.022	0.039	0.064	1.003
$\sigma_{\theta.fs}\sigma_{\theta.sl}$ (2003)	0.014	0.034	0.059	1.000
$\sigma^2_{\theta.fs}$ (2004)	0.041	0.076	0.124	1.003
$\sigma^2_{\theta.sl}$ (2004)	0.017	0.034	0.056	1.003
$\sigma_{\theta.fs}\sigma_{\theta.sl}$ (2004)	0.014	0.035	0.062	1.000
$\sigma^2_{\theta.fs}$ (2005)	0.036	0.070	0.115	1.000
$\sigma^2_{\theta.sl}$ (2005)	0.018	0.034	0.058	1.000
$\sigma_{\theta.fs}\sigma_{\theta.sl}$ (2005)	0.007	0.026	0.051	1.001

Appendix D

Parameter estimates for selected ward models (West Midlands)

Table D.1: Model PLN

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs[1])	1.79	1.87	1.95	1.001
Intercept (sl[2])	3.78	3.83	3.88	1.001
Population (fs)	-0.05	0.18	0.39	1.001
Population (sl)	0.07	0.22	0.38	1.011
Area (fs)	0.03	0.20	0.37	1.001
Area (sl)	-0.09	0.03	0.15	1.001
Major road (fs)	-0.02	0.07	0.17	1.001
Major road (sl)	0.10	0.16	0.21	1.001
Minor road (fs)	-0.27	-0.05	0.18	1.002
Minor road (sl)	-0.02	0.12	0.27	1.001
Junction (fs)	0.01	0.14	0.29	1.004
Junction (sl)	- 0.01	0.06	0.13	1.002
Bus (fs)	-0.03	0.12	0.26	1.002
Bus (sl)	0.06	0.16	0.25	1.001
Car (driver) (fs)	-0.40	-0.25	-0.11	1.001
Car (driver)(sl)	-0.31	-0.21	-0.12	1.000
Car (passenger)(fs)	-0.17	-0.04	0.10	1.002
Car (passenger) (sl)	-0.23	-0.14	-0.05	1.000
Foot (fs)	0.08	0.21	0.33	1.008
Foot (sl)	0.13	0.21	0.28	1.002

Continued on next page...

[1] 'fs': fatal and serious accidents.

[2] 'sl': for slight accidents.

Table D.1 – Continued

Parameter	2.5%	Median	97.5%	$\hat{R}$
σ_v[1]	0.22	0.26	0.30	1.001

Table D.2: Model PLNre

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs)	1.79	1.87	1.95	1.001
Intercept (sl)	3.78	3.83	3.88	1.001
Population (fs)	-0.17	0.07	0.35	1.001
Population (sl)	0.11	0.30	0.47	1.007
Area (fs)	-0.03	0.14	0.32	1.002
Area (sl)	-0.04	0.07	0.18	1.015
Major road (fs)	0.02	0.11	0.20	1.000
Major road (sl)	0.09	0.15	0.21	1.000
Minor road (fs)	-0.27	-0.04	0.20	1.002
Minor road (sl)	-0.01	0.10	0.24	1.005
Junction (fs)	0.01	0.14	0.29	1.005
Junction (sl)	0.00	0.07	0.15	1.001
Bus (fs)	-0.18	0.01	0.20	1.019
Bus (sl)	0.13	0.25	0.38	1.020
Car (driver) (fs)	-0.35	-0.23	-0.09	1.000
Car (driver)(sl)	-0.30	-0.21	-0.12	1.000
Car (passenger)(fs)	-0.19	-0.03	0.12	1.005
Car (passenger) (sl)	-0.28	-0.18	-0.08	1.008
Foot (fs)	0.09	0.21	0.34	1.002
Foot (sl)	0.13	0.21	0.29	1.005
σ_v	0.21	0.25	0.29	1.007
Birmingham (fs)	0.00	0.57	1.17	1.050
Birmingham (sl)	-0.72	-0.33	-0.01	1.040
Coventry (fs)	0.17	0.53	0.91	1.013
Coventry (sl)	-0.29	-0.06	0.17	1.007
Sandwell (fs)	0.00	0.31	0.60	1.007
Sandwell (sl)	-0.20	-0.02	0.18	1.003
Solihull (fs)	-0.03	0.34	0.72	1.000
Solihull (sl)	-0.24	0.00	0.22	1.004
Dudley (fs)	-0.35	0.00	0.31	1.004
Dudley (sl)	-0.10	0.11	0.31	1.001

Continued on next page...

[1] σ_v: standard deviation of unstructured random effects

Table D.2 – Continued

Parameter	2.5%	Median	97.5%	$\hat{R}$
Walsall (fs)	-0.28	0.02	0.31	1.012
Walsall (sl)	-0.38	-0.20	0.01	1.008

Table D.3: Model MVCCAR.mv

Parameter	2.5%	Median	97.5%	$\hat{R}$
Intercept (fs)	1.79	1.87	1.94	1.003
Intercept (sl)	3.78	3.83	3.88	1.001
Population (fs)	-0.16	0.10	0.33	1.010
Population (sl)	0.00	0.17	0.32	1.037
Area (fs)	-0.01	0.18	0.36	1.004
Area (sl)	-0.06	0.04	0.15	1.002
Major road (fs)	0.00	0.10	0.20	1.004
Major road (sl)	0.10	0.16	0.22	1.005
Minor road (fs)	-0.27	-0.04	0.20	1.003
Minor road (sl)	0.00	0.13	0.24	1.003
Junction (fs)	0.03	0.17	0.33	1.002
Junction (sl)	0.02	0.09	0.17	1.003
Bus (fs)	-0.01	0.16	0.34	1.016
Bus (sl)	0.08	0.20	0.32	1.040
Car (driver) (fs)	-0.35	-0.20	-0.03	1.001
Car (driver)(sl)	-0.29	-0.19	-0.09	1.009
Car (passenger)(fs)	-0.23	-0.07	0.07	1.001
Car (passenger) (sl)	-0.26	-0.16	-0.07	1.000
Foot (fs)	0.09	0.21	0.33	1.002
Foot (sl)	0.13	0.21	0.28	1.009
$\sigma^2_{v.fs}$	0.01	0.05	0.11	1.002
$\sigma^2_{v.sl}$	0.01	0.04	0.07	1.014
$\sigma_{v.fs}\sigma_{v.sl}$	0.00	0.03	0.06	1.005
$\sigma^2_{\theta.fs}$ [1]	0.03	0.11	0.34	1.015
$\sigma^2_{\theta.sl}$	0.02	0.08	0.20	1.008
$\sigma_{\theta.fs}\sigma_{\theta.sl}$	0.00	0.06	0.21	1.013

[1] σ_θ: standard deviation of spatial random effects

Appendix E

Predictions using the non-CAR model

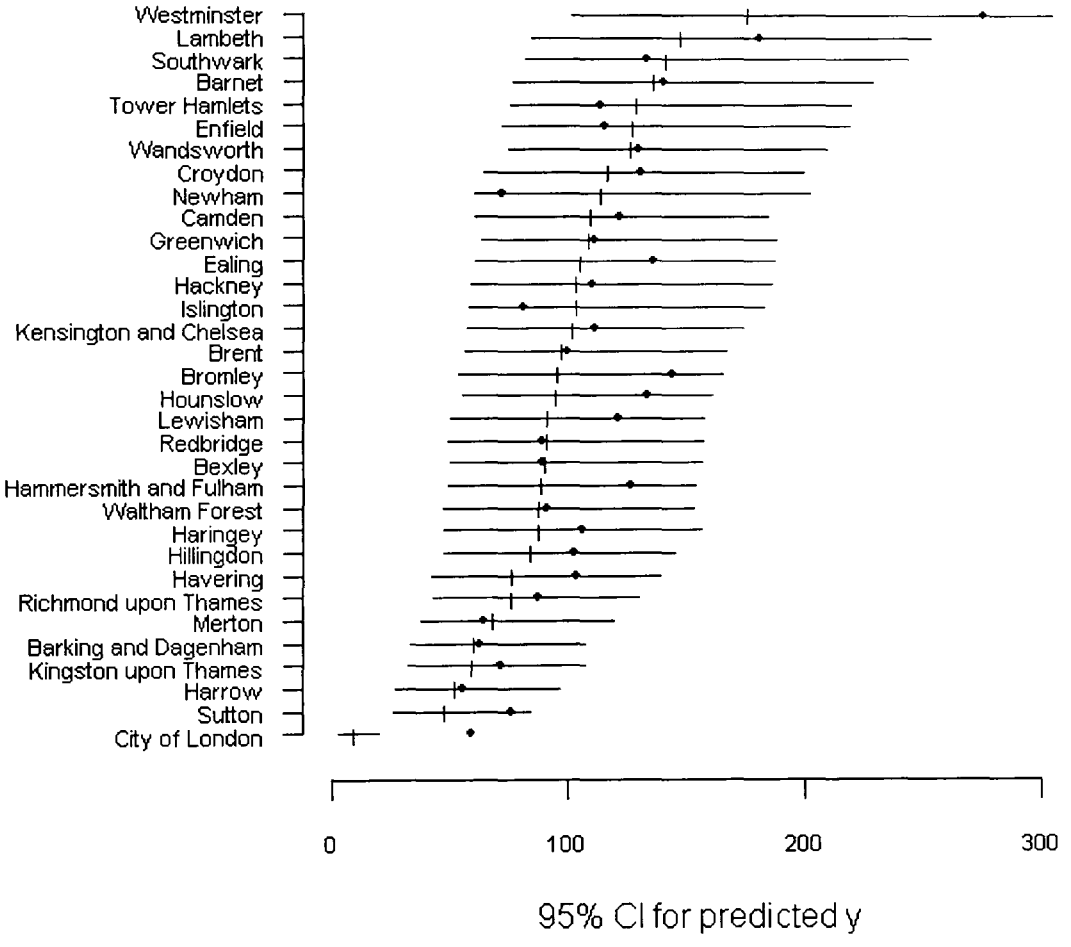

Figure E.1: Predictions for London boroughs.

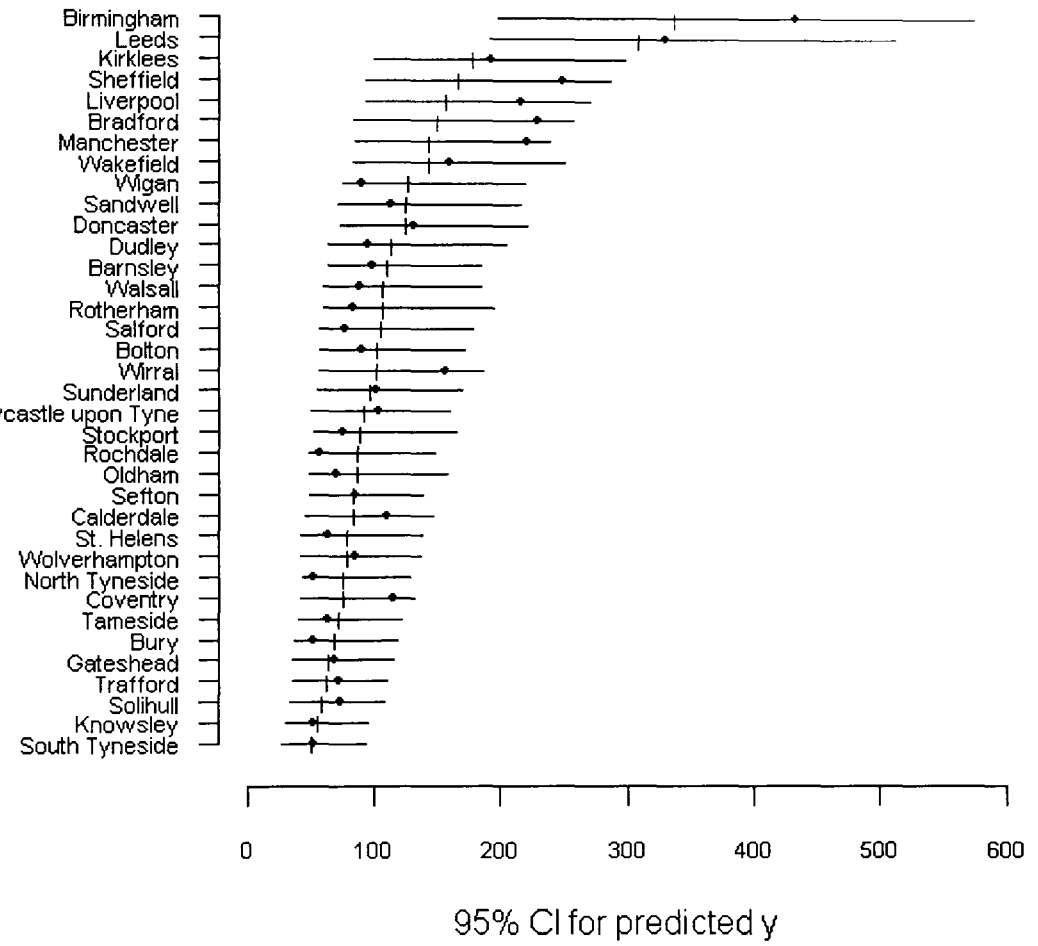

Figure E.2: Predictions for metropolitan districts.

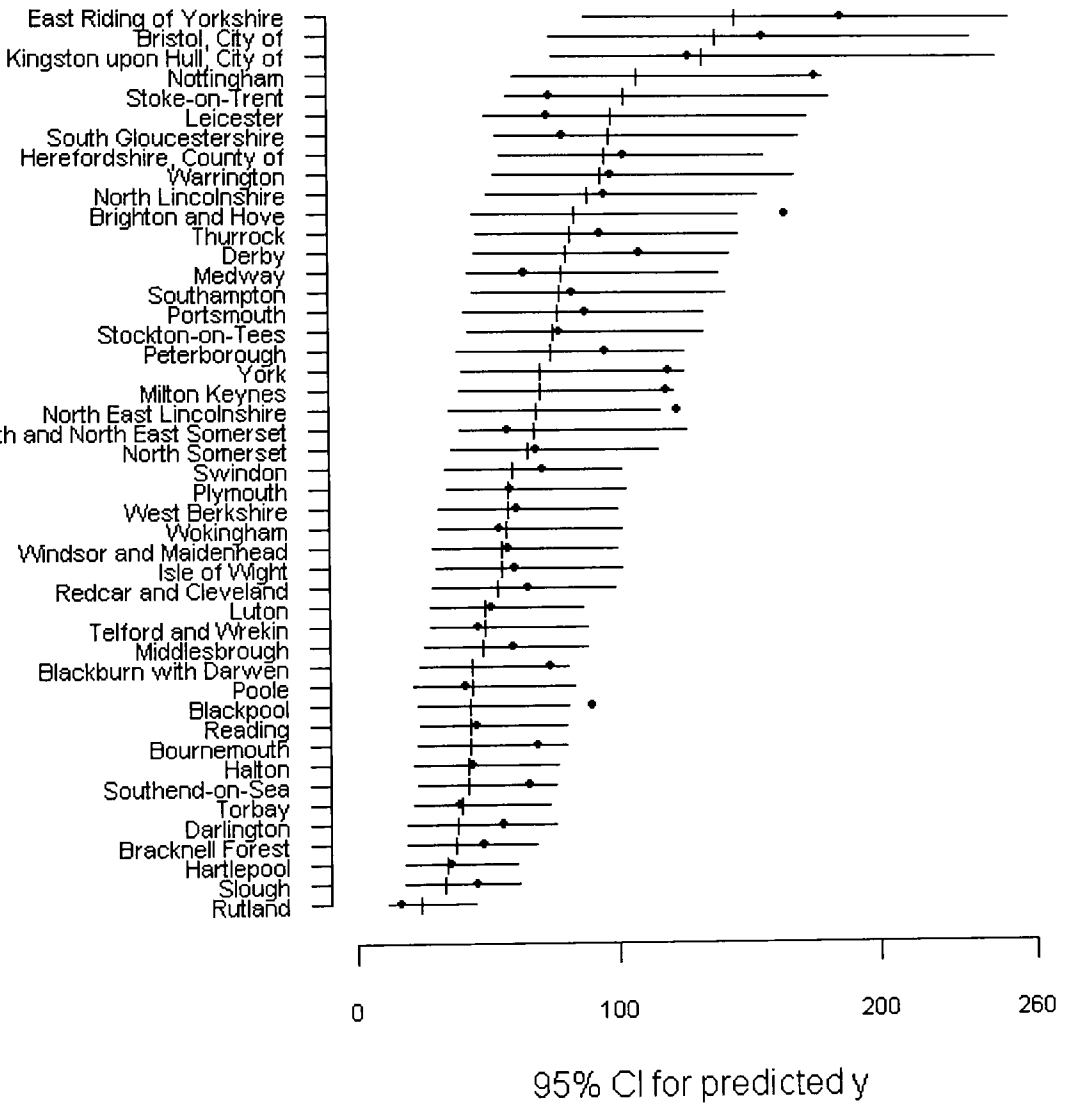

Figure E.3: Predictions for unitary authorities.

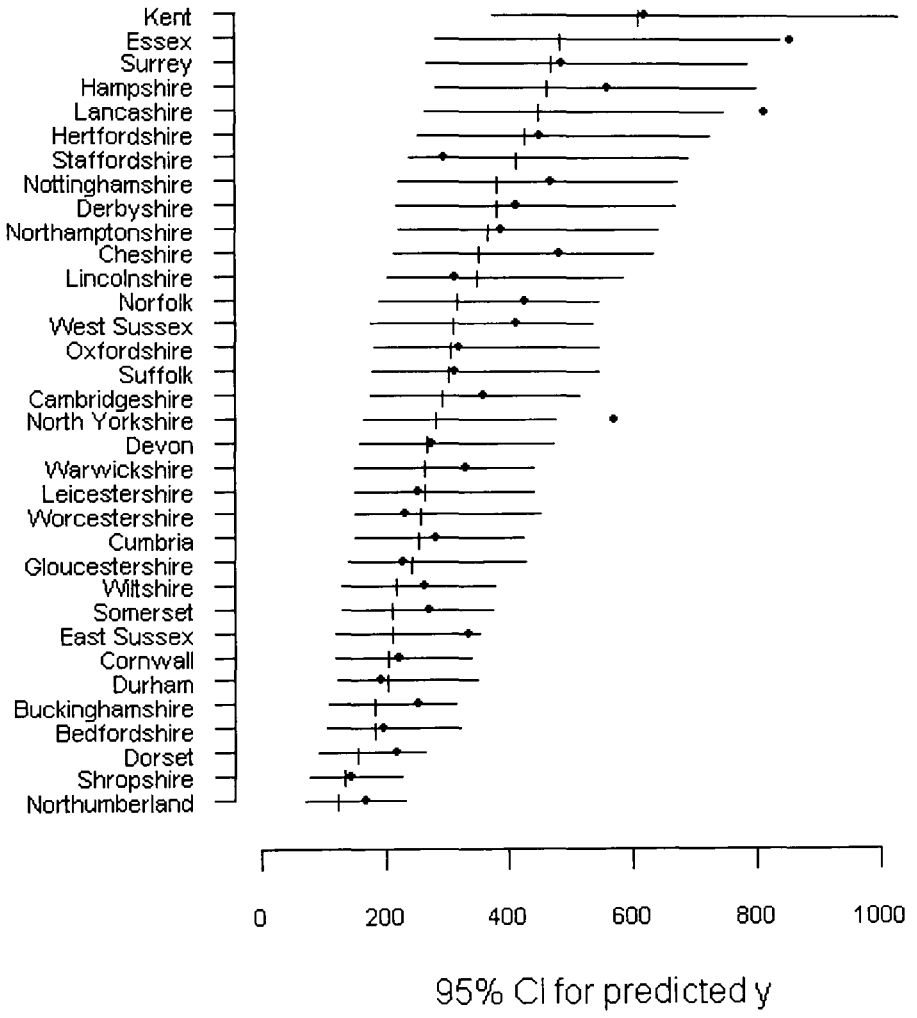

Figure E.4: Predictions for other local authorities.

Appendix F

WinBUGS codes for selected models

F.1 Areal models (1983 - 1986)

```
model { # Poisson regression model with log-normal random effects
# Accidents on built-up A-roads disaggregated by severity from 1983 to 1986
for(t in 1:4){
for (i in 1 : 108) {
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <- beta0 + beta1*AREA[i,t] + beta2 * POP[i,t] + beta3 *VEH[i,t]
+ beta4 *ROAD[i,t]+beta5 *TRAFFIC[i,t] + v[i,t]
v[i,t] ~ dnorm(0,tau.v) #unstructured random effects
}
      }
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm( 0.0, 0.0001)
beta2 ~ dnorm( 0.0, 0.0001)
beta3 ~ dnorm( 0.0, 0.0001)
beta4 ~ dnorm( 0.0, 0.0001)
beta5 ~ dnorm( 0.0, 0.0001)
tau.v ~ dgamma(0.01,0.01)
sigma.v<- 1/sqrt(tau.v)
}
      model {# Poisson regression model with metropolitan county effects
for(t in 1:4){
for (i in 1 : 108) {
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <-beta0 + alpha1*LON[i,t] + alpha2*MAN[i,t] + alpha3*TYNE[i,t] + alpha4*WYORK[i,t] + alpha5*SYORK[i,t] +
alpha6*MER[i,t] + alpha7*MID[i,t] +
beta1*AREA[i,t] + beta2 * POP[i,t] + beta3 *VEH[i,t]
+beta4 *ROAD[i,t]+beta5 *TRAFFIC[i,t]
```

```
}
}
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm( 0.0, 0.0001)
beta2 ~ dnorm( 0.0, 0.0001)
beta3 ~ dnorm( 0.0, 0.0001)
beta4 ~ dnorm( 0.0, 0.0001)
beta5 ~ dnorm( 0.0, 0.0001)
alpha1 ~ dnorm(0.0, 0.0001)
alpha2 ~ dnorm(0.0, 0.0001)
alpha3 ~ dnorm(0.0, 0.0001)
alpha4 ~ dnorm(0.0, 0.0001)
alpha5 ~ dnorm(0.0, 0.0001)
alpha6 ~ dnorm(0.0, 0.0001)
alpha7 ~ dnorm(0.0, 0.0001)
}

    model {# convolution CAR model
for(t in 1:4){
for (i in 1 : 108) {
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <-beta0 + beta1*AREA[i,t] + beta2 * POP[i,t] + beta3 *VEH[i,t]+
beta4 *ROAD[i,t]+beta5 *TRAFFIC[i,t] + v[i,t] + theta[t,i]
v[i,t] ~ dnorm(0,tau.v)
}
# CAR prior for spatial random effects:
theta[t,1:108]~ car.normal(adj[], weight[], num[], tau.theta[t])
}
beta0~dflat()
beta1 ~ dnorm( 0.0, 0.0001)
beta2 ~ dnorm( 0.0, 0.0001)
beta3 ~ dnorm( 0.0, 0.0001)
beta4 ~ dnorm( 0.0, 0.0001)
beta5 ~ dnorm( 0.0, 0.0001)
for (t in 1:4){
tau.theta[t] ~ dgamma(0.5, 0.0005)
sigma.theta[t] <- sqrt(1 / tau.theta[t]) #calculate variance
}
tau.v ~ dgamma(0.5, 0.0005)
sigma.v <- sqrt(1 / tau.v) }
```

F.2 Areal models (2001 - 2005)

```
model {#Poisson regression model with log-normal random effects and a linear time trend
for(t in 1:5){
for (i in 1 : 149) {
y1[i,t] ~ dpois(lambda1[i,t])
y2[i,t] ~ dpois(lambda2[i,t])
log(lambda1[i,t]) <-beta0[1] + delta[1]*(t-1) +
beta1[1]*AREA[i] + beta2[1]* POP[i,t] + beta3[1] *AROAD[i] +
beta4[1]* BROAD[i] + beta5[1] * MROAD[i] + beta6[1] * TROTHER[i,t] +
beta7[1]* TRCAR[i,t]+ beta8[1]*NODE[i] + v[i,t]
log(lambda2[i,t]) <-beta0[2] + delta[2]*(t-1) +
beta1[2]*AREA[i] + beta2[2]* POP[i,t] + beta3[2] *AROAD[i] +
beta4[2]* BROAD[i] + beta5[2] * MROAD[i] + beta6[2] * TROTHER[i,t] +
beta7[2]* TRCAR[i,t]+ beta8[2]*NODE[i] + v[i,t]
v[i,t] ~ dnorm(0,tau.v)
}
}
tau.v ~ dgamma(0.5,0.0005)
sigma.v <- 1/tau.v
for (i in 1:2){
beta0[i] ~ dnorm(0,0.0001)
beta1[i] ~ dnorm( 0.0, 0.0001)
beta2[i] ~ dnorm( 0.0, 0.0001)
beta3[i] ~ dnorm( 0.0, 0.0001)
beta4[i] ~ dnorm( 0.0, 0.0001)
beta5[i] ~ dnorm( 0.0, 0.0001)
beta6[i] ~ dnorm( 0.0, 0.0001)
beta7[i] ~ dnorm( 0.0, 0.0001)
beta8[i] ~ dnorm( 0.0, 0.0001)
delta[i] ~ dnorm(0,0.0001)
}}
```

```
    model {# CAR model with temporal effects
for(t in 1:5){
for (i in 1 : 149) {
y1[i,t] ~ dpois(lambda1[i,t])
y2[i,t] ~ dpois(lambda2[i,t])
log(lambda1[i,t]) <-beta0[1] + delta[1]*(t-1) + alpha[1]*unitary[i]+
beta1[1]*AREA[i] + beta2[1]* POP[i,t] + beta3[1] *AROAD[i] +
beta4[1]* BROAD[i] + beta5[1] * MROAD[i] + beta6[1] * TROTHER[i,t] +
beta7[1]* TRCAR[i,t]+ beta8[1]*NODE[i]+ theta.fs[t,i] + u.fs[i,t]
log(lambda2[i,t]) <-beta0[2] + delta[2]*(t-1) +alpha[2]*unitary[i]+
beta1[2]*AREA[i] + beta2[2]* POP[i,t] + beta3[2] *AROAD[i] +
beta4[2]* BROAD[i] + beta5[2] * MROAD[i] + beta6[2] * TROTHER[i,t] +
beta7[2]* TRCAR[i,t]+ beta8[2]*NODE[i] + theta.sl[t,i] + u.sl[i,t]
#assume a first order autoregressive prior for temporal effects
u.fs[i,t] ~ dnorm(v.fs[t],tau.u.fs)
u.sl[i,t] ~ dnorm(v.sl[t],tau.u.sl)
}
theta.fs[t,1:149] ~ car.normal(adj[], weight[], num[], tau.theta.fs[t])
theta.sl[t,1:149] ~ car.normal(adj[], weight[], num[], tau.theta.sl[t])
} # specify first order autoregressive prior
for (t in 2:5) {
v.fs[t]<-rho[1]*v.fs[t-1]
v.sl[t]<-rho[2]*v.sl[t-1]
}
v.fs[1] <- k1
v.sl[1] <- k2
k1 ~ dnorm(0,tau.u.fs)
k2 ~ dnorm(0,tau.u.sl)
rho[1] ~ dunif(-1,1)
rho[2] ~ dunif(-1,1)
tau.u.fs ~ dgamma(0.5,0.0005)
sigma.u.fs <- 1/tau.u.fs
tau.u.sl ~ dgamma(0.5,0.0005)
sigma.u.sl <- 1/tau.u.sl
for (t in 1:5){
tau.theta.fs[t]~ dgamma(0.5, 0.0005)
tau.theta.sl[t] ~ dgamma(0.5, 0.0005)
sigma.theta.fs[t]<-1/tau.theta.fs[t]
sigma.theta.sl[t]<-1/tau.theta.sl[t]
}
for (i in 1:2){
beta0[i] ~ dflat()
delta[i]~dnorm(0,0.0001)
alpha[i] ~ dnorm( 0.0, 0.0001)
```

beta1[i] ~ dnorm(0.0, 0.0001)

beta2[i] ~ dnorm(0.0, 0.0001)

beta3[i] ~ dnorm(0.0, 0.0001)

beta4[i] ~ dnorm(0.0, 0.0001)

beta5[i] ~ dnorm(0.0, 0.0001)

beta6[i] ~ dnorm(0.0, 0.0001)

beta7[i] ~ dnorm(0.0, 0.0001)

beta8[i] ~ dnorm(0.0, 0.0001)}}

```
model {# multivariate CAR model with temporal effects
for(t in 1:5){
for (i in 1 : 149) {
y1[i,t] ~ dpois(lambda1[i,t])
y2[i,t] ~ dpois(lambda2[i,t])
log(lambda1[i,t]) <-beta0[1] + delta[1]*(t-1) + alpha[1]*unitary[i]+
beta1[1]*AREA[i] + beta2[1]* POP[i,t] + beta3[1] *AROAD[i] +
beta4[1]* BROAD[i] + beta5[1] * MROAD[i] + beta6[1] * TROTHER[i,t] +
beta7[1]* TRCAR[i,t]+ beta8[1]*NODE[i]+ xi[i,t,1] + theta[t,1,i] + u.fs[i,t]
log(lambda2[i,t]) <-beta0[2] + delta[2]*(t-1) +alpha[2]*unitary[i]+
beta1[2]*AREA[i] + beta2[2]* POP[i,t] + beta3[2] *AROAD[i] +
beta4[2]* BROAD[i] + beta5[2] * MROAD[i] + beta6[2] * TROTHER[i,t] +
beta7[2]* TRCAR[i,t]+ beta8[2]*NODE[i] + xi[i,t,2] + theta[t,2,i] + u.sl[i,t]
xi[i,t,1:2] ~ dmnorm(zero[], tau.xi[,]) #specify a multivariate normal prior
u.fs[i,t] ~ dnorm(v.fs[t],tau.u.fs)
u.sl[i,t] ~ dnorm(v.sl[t],tau.u.sl)
}
}
# use a wishart distribution as the prior for the inverse variance-covariance matrix
tau.xi[1:2,1:2] ~ dwish(Q[,],2)
#calculate the variance-covariane matrix
sigma.xi[1:2, 1:2] <- inverse(tau.xi[ , ])
Q[1,1]<-0.1 #prior for Q
Q[2,2]<-0.1
Q[1,2] <- 0
Q[2,1]<- 0
# specify spatial prior for each year
theta[1,1:2, 1:149] ~ mv.car(adj[], weight[], num[], tau.spatial.1[ ,])
theta[2,1:2, 1:149] ~ mv.car(adj[], weight[], num[], tau.spatial.2[ ,])
theta[3,1:2, 1:149] ~ mv.car(adj[], weight[], num[], tau.spatial.3[ ,])
theta[4,1:2, 1:149] ~ mv.car(adj[], weight[], num[], tau.spatial.4[ ,])
theta[5,1:2, 1:149] ~ mv.car(adj[], weight[], num[], tau.spatial.5[ ,])
# specify prior for the inverse variance-covariance matrix for the spatial effects
tau.spatial.1[1:2,1:2] ~ dwish(R[ , ], 2)
tau.spatial.2[1:2,1:2] ~ dwish(R[ , ], 2)
tau.spatial.3[1:2,1:2] ~ dwish(R[ , ], 2)
tau.spatial.4[1:2,1:2] ~ dwish(R[ , ], 2)
tau.spatial.5[1:2,1:2] ~ dwish(R[ , ], 2)
R[1,1]<-0.1
R[2,2]<-0.1
R[1,2] <- 0
R[2,1]<- 0
# calculate the variance-covariance matrix
sigma.spatial.1[1:2, 1:2] <- inverse(tau.spatial.1[,])
```

```
sigma.spatial.2[1:2, 1:2] <- inverse(tau.spatial.2[,])
sigma.spatial.3[1:2, 1:2] <- inverse(tau.spatial.3[,])
sigma.spatial.4[1:2, 1:2] <- inverse(tau.spatial.4[,])
sigma.spatial.5[1:2, 1:2] <- inverse(tau.spatial.5[,])
tau.u.fs ~ dgamma(0.5,0.0005)
sigma.u.fs <- 1/tau.u.fs
tau.u.sl ~ dgamma(0.5,0.0005)
sigma.u.sl <- 1/tau.u.sl
for (t in 2:5) {
v.fs[t]<-rho[1]*v.fs[t-1]
v.sl[t]<-rho[2]*v.sl[t-1]
}
v.fs[1] <- k1
v.sl[1] <- k2
k1 ~ dnorm(0,tau.u.fs)
k2 ~ dnorm(0,tau.u.sl)
rho[1] ~ dunif(-1,1)
rho[2] ~ dunif(-1,1)
for (i in 1:2){
beta0[i] ~ dflat()
delta[i]~dnorm(0,0.0001)
alpha[i] ~ dnorm( 0.0, 0.0001)
beta1[i] ~ dnorm( 0.0, 0.0001)
beta2[i] ~ dnorm( 0.0, 0.0001)
beta3[i] ~ dnorm( 0.0, 0.0001)
beta4[i] ~ dnorm( 0.0, 0.0001)
beta5[i] ~ dnorm( 0.0, 0.0001)
beta6[i] ~ dnorm( 0.0, 0.0001)
beta7[i] ~ dnorm( 0.0, 0.0001)
beta8[i] ~ dnorm( 0.0, 0.0001)
}
}
```

F.3 Ward models (West Midlands)

```
model
{ # Poisson regression model with log-normal random effects
      for (i in 1 :162)
y1[i] ~ dpois(lambda1[i]) # fatal and serious accidents
y2[i] ~ dpois(lambda2[i]) # slight accidents

log(lambda1[i]) <- beta0[1] + beta1[1]*pop1[i] + beta2[1]*area1[i] +
beta3[1]*major1[i] + beta4[1]*minor1[i] + beta5[1]*node1[i] + beta6[1]*bus1[i]+
beta7[1]*car11[i]+ beta8[1]*car21[i]+ beta9[1]*foot1[i] + v[i,1]
log(lambda2[i]) <- beta0[2] + beta1[2]*pop1[i] + beta2[2]*area1[i] +
beta3[2]*major1[i] + beta4[2]*minor1[i] + beta5[2]*node1[i] + beta6[2]*bus1[i]+
beta7[2]*car11[i]+ beta8[2]*car21[i]+ beta9[2]*foot1[i] + v[i,2]

      v[i,1] ~ dnorm (0,tau.v)
v[i,2] ~ dnorm (0,tau.v)
}
tau.v~dgamma(0.01,0.01)
sigma.v<-1/sqrt(tau.v)
for (i in 1:2) {
beta0[i]~dnorm(0.0,0.0001)
beta1[i] ~ dnorm( 0.0, 0.0001)
beta2[i] ~ dnorm( 0.0, 0.0001)
beta3[i] ~ dnorm( 0.0, 0.0001)
beta4[i] ~ dnorm( 0.0, 0.0001)
beta5[i] ~ dnorm( 0.0, 0.0001)
beta6[i] ~ dnorm( 0.0, 0.0001)
beta7[i] ~ dnorm( 0.0, 0.0001)
beta8[i] ~ dnorm( 0.0, 0.0001)
beta9[i] ~ dnorm( 0.0, 0.0001)
}
}

      model
{ # Poisson regression model with metropolitan county effects
      for (i in 1 :162)
y1[i] ~ dpois(lambda1[i]) # fatal and serious accidents
y2[i] ~ dpois(lambda2[i]) # slight accidents
log(lambda1[i]) <- beta0[1] +
alpha1[1]*birmingham[i] + alpha2[1]*coventry[i] + alpha3[1]*sandwell[i] +
alpha4[1]*solihull[i]+ alpha5[1]*dudley[i] + alpha6[1]*walsall[i]+
beta1[1]*pop1[i] + beta2[1]*area1[i] + beta3[1]*major1[i] + beta4[1]*minor1[i] +
beta5[1]*node1[i] + beta6[1]*bus1[i]+ beta7[1]*car11[i]+ beta8[1]*car21[i]+
beta9[1]*foot1[i] + v[i,1]
```

```
log(lambda2[i]) <- beta0[2] +
alpha1[2]*birmingham[i] + alpha2[2]*coventry[i] + alpha3[2]*sandwell[i] +
alpha4[2]*solihull[i]+ alpha5[2]*dudley[i] +alpha6[2]*walsall[i]+
beta1[2]*pop1[i] + beta2[2]*area1[i] + beta3[2]*major1[i] + beta4[2]*minor1[i] +
beta5[2]*node1[i] + beta6[2]*bus1[i]+ beta7[2]*car11[i]+ beta8[2]*car21[i]+
beta9[2]*foot1[i] + v[i,2]

    }

    for (i in 1:2) {
alpha1[i] ~ dnorm( 0.0, 0.0001)
alpha2[i] ~ dnorm( 0.0, 0.0001)
alpha3[i] ~ dnorm( 0.0, 0.0001)
alpha4[i] ~ dnorm( 0.0, 0.0001)
alpha5[i] ~ dnorm( 0.0, 0.0001)
alpha6[i] ~ dnorm( 0.0, 0.0001)
beta0[i]~dnorm(0.0,0.0001)
beta1[i] ~ dnorm( 0.0, 0.0001)
beta2[i] ~ dnorm( 0.0, 0.0001)
beta3[i] ~ dnorm( 0.0, 0.0001)
beta4[i] ~ dnorm( 0.0, 0.0001)
beta5[i] ~ dnorm( 0.0, 0.0001)
beta6[i] ~ dnorm( 0.0, 0.0001)
beta7[i] ~ dnorm( 0.0, 0.0001)
beta8[i] ~ dnorm( 0.0, 0.0001)
beta9[i] ~ dnorm( 0.0, 0.0001)
}
}
```

```
model { # multivariate CAR models

for (i in 1 :162) {
y1[i] ~ dpois(lambda1[i])
y2[i] ~ dpois(lambda2[i])
log(lambda1[i]) <- beta0[1] +
beta1[1]*pop1[i] + beta2[1]*area1[i] + beta3[1]*major1[i] + beta4[1]*minor1[i] +
beta5[1]*node1[i] + beta6[1]*bus1[i]+ beta7[1]*car11[i]+ beta8[1]*car21[i]+
beta9[1]*foot1[i] + v[i,1] + theta[1,i]
log(lambda2[i]) <- beta0[2] +
beta1[2]*pop1[i] + beta2[2]*area1[i] + beta3[2]*major1[i] + beta4[2]*minor1[i] +
beta5[2]*node1[i] + beta6[2]*bus1[i]+ beta7[2]*car11[i]+ beta8[2]*car21[i]+
beta9[2]*foot1[i] + v[i,2] + theta[2,i]
v[i,1:2] ~ dmnorm(zero[], tau.v[,]) #multivariate normal prior
}
# use a wishart distribution as the prior of the inverse variance-covariance matrix
tau.v[1:2,1:2] ~ dwish(Q[,],2)
sigma.v[1:2, 1:2] <- inverse(tau.v[ , ]) # calculate the variance-covariance matrix for v
Q[1,1]<-0.1 # initial values for Q
Q[2,2]<-0.1
Q[1,2] <- 0
Q[2,1]<- 0
# multivariate CAR prior
theta[1:2, 1:162] ~ mv.car(adj[], weight[], num[], tau.theta[ ,])
tau.theta[1:2,1:2] ~ dwish(R[ , ], 2)
sigma.theta[1:2, 1:2] <- inverse(tau.theta[,]) # calculate the variance-covariance matrix for theta
R[1,1]<-0.1 # initial values for R
R[2,2]<-0.1
R[1,2] <- 0
R[2,1]<- 0
for (i in 1:2) {
beta0[i]~dflat()
beta1[i] ~ dnorm( 0.0, 0.0001)
beta2[i] ~ dnorm( 0.0, 0.0001)
beta3[i] ~ dnorm( 0.0, 0.0001)
beta4[i] ~ dnorm( 0.0, 0.0001)
beta5[i] ~ dnorm( 0.0, 0.0001)
beta6[i] ~ dnorm( 0.0, 0.0001)
beta7[i] ~ dnorm( 0.0, 0.0001)
beta8[i] ~ dnorm( 0.0, 0.0001)
beta9[i] ~ dnorm( 0.0, 0.0001)
}
    }
```

F.4 Link models for the M1

```
model{# Poisson regression model with log-normal random effects and a linear time trend
for (i in 1:59){
for (t in 1:7){
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <- beta0 + delta*(t-1) +
beta1 * aadf[i,t] + beta2 * length[i] + v[i,t]
v[i,t] ~ dnorm(0,tau.v)
}}
tau.v ~ dgamma(0.01,0.01)
sigma.v <- 1/sqrt(tau.v)
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
delta ~ dnorm(0,0.0001)

    }

model{ # convolution CAR model with a liner time trend
for (t in 1:7){
for (i in 1:59) {
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <- beta0 + delta*(t-1)+
beta1 * aadf[i,t] + beta2 * length[i] + v[i,t] + theta[i]
v[i,t] ~ dnorm(0,tau.v)
} }
# assume spatial effects to be fixed over time
theta[1:59]~ car.normal(adj[], weight[], num[], tau.theta)
tau.theta ~ dgamma(0.5, 0.0005)
sigma.theta <- 1/sqrt(tau.theta)
beta0 ~ dflat()
tau.v ~ dgamma(0.5,0.0005)
sigma.v <-1/sqrt(tau.v)
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
delta ~ dnorm(0,0.0001)

    }
```

F.5 Junction models for Coventry

```
model{ #convolution CAR model
for(i in 1:55){
y[i]~dpois(lambda[i])
log(lambda[i])<-beta0 + theta[i] + v[i]
v[i] ~ dnorm(0,tau.v)
}
theta[1:55] ~ car.normal(adj[], weight[], num[], tau.theta)
beta0 ~ dflat()
tau.theta ~ dgamma(0.5,0.0005)
sigma.theta <- sqrt(1/tau.theta)
tau.v ~ dgamma(0.5,0.0005)
sigma.v<-sqrt(1/tau.v)
}
```

```
model{# proper CAR model
for(i in 1:55){
y[i]~dpois(lambda[i])
log(lambda[i])<- beta0 + theta[i] + v[i]
# m[i] the diagonal element of the conditional variance matrix, assumed to be inversely proportional to the number of neighbours for
junction i
m[i]<-1/num[i]
M[i] <- alpha #the mean for junction i
v[i]~ dnorm(0,tau.v)
}
theta[1:55]~car.proper(M[], weight[], adj[], num[], m[], tau.theta, gamma)
alpha ~ dnorm(0, 0.0001)
tau.theta ~ dgamma(0.5, 0.0005) # prior on precision
sigma.theta <- 1/tau.theta # variance
gamma ~ dunif(gamma.min, gamma.max) # overall degree of spatial dependence
gamma.min <- min.bound(C[], adj[], num[], m[]) #lower bound
gamma.max <- max.bound(C[], adj[], num[], m[]) # upper bound
tau.v~ dgamma(0.01,0.001)
sigma.v<-1/sqrt(tau.u)
beta0 ~ dnorm(0,0.0001)
}
```

F.6 Areal model for prediction

```
model { # CAR model
for(t in 1:6){
for (i in 1 : 149) {
y[i,t] ~ dpois(lambda[i,t])
log(lambda[i,t]) <-beta0 + delta*(t-1) + alpha*unitary[i] +
beta1*AREA[i] + beta2* POP[i,t] + beta3 *AROAD[i] + beta4* BROAD[i] +
beta5 * MROAD[i] +beta6 * TROTHER[i,t] + beta7* TRCAR[i,t]+
beta8*NODE[i]+ theta.fs[t,i] + u.fs[i,t]
#assume a first order autoregressive prior for temporal effects
u[i,t] ~ dnorm(v[t],tau.u)
}
}
for (t in 2:6){
v[t]<-rho*v[t-1]
}
v[1]<-k
k ~ dnorm(0,tau.u)
rho ~ dunif(-1,1)
tau.u ~ dgamma(0.5,0.0005)
sigma.u <- 1/tau.u
for(i in 1:149){
for(t in 1:3){
#assume spatial effects are fixed in the first period
theta.fs[t,i]<-theta.fs.1[i]
}
for(t in 4:6){
#assume spatial effects are fixed in second first period
theta.fs[t,i]<-theta.fs.2[i]
}
}
theta.fs.1[1:149] ~ car.normal(adj[], weight[], num[], tau.fs.1)
theta.fs.2[1:149] ~ car.normal(adj[], weight[], num[], tau.fs.2)
tau.fs.1~ dgamma(0.5, 0.0005)
tau.fs.2 ~ dgamma(0.5, 0.0005)
beta0~ dflat()
beta1 ~ dnorm( 0.0, 0.0001)
beta2 ~ dnorm( 0.0, 0.0001)
beta3 ~ dnorm( 0.0, 0.0001)
beta4 ~ dnorm( 0.0, 0.0001)
beta5 ~ dnorm( 0.0, 0.0001)
beta6 ~ dnorm( 0.0, 0.0001)
beta7 ~ dnorm( 0.0, 0.0001)
beta8 ~ dnorm( 0.0, 0.0001)
```

```
alpha ~ dnorm( 0.0, 0.0001)
delta ~ dnorm( 0.0, 0.0001)
#save prediction results for 2006
for (i in 1:149){
pred[i]<-fs[i,6] #predicted y
pred.lambda[i]<-lambda[i,6] #Poisson mean
}}
```

References

Abbess, C., Jarrett, D. and Wright, C. C. (1981), 'Accidents at blackspots: Estimating the effectiveness of remedial treatment with special reference to the regression-to-mean effect', *Traffic Engineering and Control* **22**, 535–542. 22

Abdel-Aty, M. A. and Radwan, A. E. (2000), 'Modeling traffic accident occurrence and involvement', *Accid. Anal. and Prev.* **32**, 633–642. 21

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**, 716–723. 36

Albert, J. and Chib, S. (1996), 'Bayesian residual analysis for binary response regression models', *Biometrika* **82**, 747–759. 36

Anderson, T. (2007), 'Comparison of spatial methods for measuring road accident 'hotspots': a case study of london', *Journal of Maps* pp. 55–63.
URL: *http://www.journalofmaps.com/viewMap.php?mid=72* 28

ATKINS (2006), *M1 high occupance vehicle lane pilot 'before' monitoring study.*
URL: *http://www.ha-research.gov.uk/projects/* 175, 176

Bailey, T. C. and Hewson, P. J. (2004), 'Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model', *J. R. Stat. Soc.* **167**, 501–517. 4, 6, 39, 79

Banerjee, S., Carlin, B. and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall. 34, 36, 37, 41, 42, 43, 44, 45

Barker, J., Farmer, S. and Nicholls, D. (1998), Injury accidents on rural single-carriageway roads, 1994-95: an analysis of stats19 data, Technical Report TRL304, Transport Research Laboratory. 13

Berhanu, G. (2004), 'Models relating traffic safety with road environment and traffic flows on arterial roads in addis ababa', *Accid. Anal. and Prev.* **36**, 697–704. 18, 21

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M. (1995), 'Bayesian analysis of space-time variation in disease risk', *Statistics in Medicine* **14**, 2433–2443. 50

Besag, J. and Kooperberg, C. (1995), 'On conditional and intrinsic autoregressions', *Biometrika* **82**, 733–746. 59

Besag, J. and Mollié, A. (1989), 'Bayesian mapping of mortality rates', *Bulletin of the International Statistical Institute* **53**, 127–128. 46

Besag, J., York, J. and Mollie, A. (1991), 'Bayesian image restoration with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**, 1–59. 46, 50

Best, N. G., Ickstadt, K., Wolpert, R. L. and Briggs, D. J. (2000), Combining models of health and exposure data: the SAVIAH study, *in* 'Spatial epidemiology: methods and applications', Oxford: Oxford University Press. 42, 47, 48

Best, N., Richardson, S. and Thomson, A. (2005), 'A comparison of Bayesian spatial models for disease mapping.', *Stat Methods Med Res* **14**, 35–59. 7, 30, 42

Bivand, R. (2004), *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.2-22.
URL: *http://cran.r-project.org/src/contrib/Descriptions/spdep.html* 52, 61, 71

Boyle, A. J. and Wright, C. C. (1984), 'Accident 'migration' after remedial treatment at accident blackspots', *Traffic Engineering and Control* **25**, 260–267. 29

Bradley, E. (2005), Modern science and the Bayesian-Frequentist controversy, Technical Report 2005-19B/233, Dept. of Statistics, Stanford Univ. 32

Brooks, S. P. and Gelman, A. (1998), 'Alternative methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**. 35

Carlin, B. and Banerjee, S. (2003), Hierarchical multivariate car models for spatially correlated survival data, *in* 'Bayesian Statistics 7', Oxford: Oxford University Press. 49

Carlin, B. P. and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall. 33, 35

Cliff, A. D. and Ord, J. K. (1981), *Quantitative geography: a British view*, Routledge & Kegan Paul, London, chapter Spatial and temporal analysis: autocorrelation in space and time, pp. 104–110. 101

Department for Transport (2001), *Tomorrow's Roads – Safer for Everyone.*
URL: *http://www.dft.gov.uk/pgr/roadsafety/* 1, 115

Department for Transport (2004), *How national traffic estimates are made.*
URL: *http://www.dft-matrix.net/forms/estimates.aspx* 77

Department for Transport (2005), *Road Accident Data - GB Variables and Values and Export Record Layouts*. 8, 14

Department for Transport (2006*a*), *Annual average daily traffic flows.*
URL: *http://www.dft-matrix.net* 87

Department for Transport (2006*b*), *Road accidents casualties: a comparison of STATS19 with Hospital Episode Statistics.*
URL: *http://www.dft.gov.uk/pgr/roadsafety/research/rsrr/theme5/* 184

Department for Transport (2006*c*), *Road Casualties Great Britain: 2005*. London: The Stationary Office. 1

Department for Transport (2007*a*), *Highways economics note No. 1 2005 valuation of the benefits of prevention of road accidents and casualties.*

URL: *http://www.dft.gov.uk/pgr/roadsafety/ea/pdfeconnote105* 1

Department for Transport (2007*b*), *Road Accident Statistics Branch, Road Accident Data, 2001 – 2006 (computer file).* Colchester, Essex: UK Data Archive. 82, 161

Department for Transport (2007*c*), *Road Casualties Great Britain 2006.* London: The Stationary Office. 184

Department for Transport (2007*d*), *Transport statistics.*

URL: *http://www.dft.gov.uk/pgr/statistics/* 83, 161

Department of Transport (1986), *Accident Investigation Manual.* London: DOT. 2

EDINA (2007), *UKBORDERS*, EDINA.

URL: *http://borders.edina.ac.uk/html/* 81, 82

Flahaut, B., Mouchart, M., Martin, E. S. and Thomas, I. (2003), 'The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach', *Accid. Anal. and Prev.* **35**, 991–1004. 13, 28

Ge, L. and Zhang, T. (2006), 'Loglinear residual tests of moran's i autocorrelation and their applications to kentucky breast cancer data', *Geographic Analysis* p. to appear. 53

Gelfand, A. and Vounatsou, P. (2003), 'Proper multivariate conditional models for spatial data analysis', *Biostatistics* (4), 11–25. 49

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', **3**. 39

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2 edn, London: Chapman and Hall. 19, 22, 36, 161

Gelman, A., Goegebeur, Y., Tuerlinckx, F. and Mechelen, I. (2000), 'Diagnostic checks for discrete data regression models using posterior predictive simulations', *Applied Statistics* **49**, 247–268. 53, 73

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences (with discussion)', *Statistical Science* **7**. 35

Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. (1998), 'Generalized linear models for small area estimation', *Journal of the American Statistical Association* **96**, 273–282. 45

Ghosh, M. and Rao, J. N. K. (1995), 'Small area estimation: An appraisal', *Statistical Science* **9**, 55–76. 43

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996), Introducing markov chain monte carlo, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman and Hall, London, pp. 1–20. 34

Greibe, P. (2003), 'Accident prediction models for urban roads', *Accid. Anal. and Prev.* **35**, 273–285. 3, 13

Hall, R. D. (1986), Accidents at four-arm single carriageway urban traffic signals, Technical Report CR65, Transport Research Laboratory. 16, 18

Hauer, E. (1997), *Observational before-after Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*, Elsevier science. 3

Hauer, E., Allery, B., Kononov, J. and Griffith, M. (2004), 'How best to rank sites with promised', *Transport. Res. Record* **1897**, 48–54. 13, 168

Hauer, E., Harwood, D., Council, F. and Griffith, M. (2002), 'The empirical Bayes method for estimating safety: A tutorial', *Transportation Research Record* **1784**, 126–131. 22, 168

Hewson, P. (2005), 'Epidemiology of child pedestrian casualty rates: Can we assume spatial independence?', *Accid. Anal. and Prev.* **37**, 651–659. 59

Highways Agency (2005), *M1 Junctions 1-19 and M10 Route Management Strategy*. **URL:** *http://www.highways.gov.uk/roads/documents/* 175

Highways Agency (2006), *M1 Jct 6a to 10 Widening*.

URL: *http://www.highways.gov.uk/roads/projects/4478.aspx* 175

Highways Agency (2007), *Road projects*.

URL: *http://www.highways.gov.uk/roads/* 175

Hirst, W., Mountain, L. and Maher, M. (2005), 'Are speed enforcement cameras more effective than other speed management measures?: An evaluation of the relationship between speed and accident reductions', *Accid. Anal. and Prev.* **37**, 731–741. 13

Jarrett, D., Abbess, C. and Wright, C. (1982), Bayesian methods applied to road accident blackspot studies: Some recent progress, *in* 'Seminar on Short-term and Area-wide Evaluation of Safety Measures', Institute for Road safety Research SWOV. 22

Jarrett, D., Hillier, H. and Wright, C. C. (1989), 'Comparisons between local authority road accident rates'. on behalf of Transport and Road Research Laboratory. 2, 3, 15, 18, 26, 79, 80

Jin, X., Carlin, B. and Banerjee, S. (2005), 'Generalized hierarchical multivariate car models for areal data', *Biostatistics* (4), 950–961. 49, 70

Jones, A., Langford, I. and Bentham, G. (1996), 'The application of k-function analysis to the geographical distribution of road traffic accident outcomes in norfolk, england', *Soc. Sci. Med.* **42**, 879–885. 28

Joseph, M. H. (2007), *Negative Binomial Regression*, Cambridge University Press. 20

Jovanis, P. P. and Chang, H. L. (1986), 'Modelling the relationship of accidents to miles travelled', *Transport. Res. Record* **1068**, 42–51. 18

Kelsall, J. and Wakefield, J. (1999), Discussion of 'Bayesian models for spatially correlated disease and exposure data' by Best et al., *in* J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, 'Bayesian Statistics 6', Chapman and Hall, London, p. 151. 60

Kim, H., Sun, D. and Tsutakawa, R. (2001), 'A bivariate Bayes methods for improving the estimates of mortality rates with a twofold conditonal autoregressive model', *Journal of the American Statistical Association* **96**, 1506–1521. 49

Knorr-Held, L. (2000), 'Bayesian modelling of inseparable space-time variation in disease risk', *Statistics in Medicine* **19**, 2555–2567. 50, 51

Knorr-Held, L. and Besag, J. (1998), 'Modelling risk from a disease in time and space'. **URL:** *citeseer.ist.psu.edu/knorr-held97modelling.html* 50

Knorr-Held, L. and Best, N. (2001), 'a shared component model for detecting joint and selective clustering of two diseases', *Journal of the Royal Statistical Society, Series A* **164**, 73–85. 51

Layfield, R., Summersgill, I., Hall, R. and Chatterjee, K. (1996), Accidents at urban priority crossroads and staggered junctions, Technical Report TRL185, Transport Research Laboratory. 3, 16

Levine, N., Kim, K. and Nitz, L. (1995*a*), 'Spatial analysis of honolulu motor vehicle crashes. i. spatial patterns', *Accid. Anal. and Prev.* **27**, 663–674. 28

Levine, N., Kim, K. and Nitz, L. (1995*b*), 'Spatial analysis of honolulu motor vehicle crashes: Ii. zonal generators', *Accid. Anal. and Prev.* **27**, 675–685. 3, 30

Lewin-Koh, N. J. and Bivand, R. (2004), *maptools: tools for reading and handling shapefiles*. R package version 0.4-7. **URL:** *http://cran.r-project.org/src/contrib/Descriptions/maptools.html* 71

Loveday, J. and Jarrett, D. (1991), Spatial autocorrelation and road accident 'migration', *in* 'the 23rd Annual conference of the universities transport studies group', Nottingham University. 29

Loveday, J. and Jarrett, D. (1992), *Mathematics in Transport Planning and Control*, Oxford: Clarendon Press, chapter Spatial Modelling of Road Accident Data, pp. 433–446. 29, 30, 47

MacNab, Y. C. (2003), 'A Bayesian hierarchical model for accident and injury surveillance', *Accid. Anal. and Prev.* **35**, 91–102. 6

Maher, M. (1987), 'Accident migration: A statistical explanation', *Traffic Engineering and Control* **28**, 480–483. 29

Maher, M. J. and Mountain, L. J. (1988), 'The identification of accident blackspots: A comparison of current methods', *Accid. Anal. and Prev.* **20**, 143–151. 13

Maher, M. J. and Summersgill, I. (1996), 'A comprehensive methodology for the fitting of predictive accidents models', *Accid. Anal. and Prev.* **28**, 281–296. 3, 4, 13, 16, 18, 21

Mardia, K. V. (1988), 'Multi-dimensional multivariate gaussian markov random fields with application to image processing', *Journal of Multivariate Analysis* **24**, 265–284. 49

Maycock, G. and Hall, R. D. (1984), Accidents at 4-arm roundabouts, Technical Report LR1120, Transport Research Laboratory. 16, 18

McCullagh, P. (2002), 'What is a statistical model?', *Annals of Statistics* **30**, 1225–1310. 5, 177

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2 edn, London: Chapman & Hall. 2, 17, 18, 37

Meridian^{TM} (2007), *Meridian^{TM} 2: User guide and technical specification*, Ordnance Survey.
URL: *http://digimap.edina.ac.uk/main* 61, 82, 83

Meyer, D., Zeileis, A. and Hornik, K. (2007), *vcd: Visualizing Categorical Data.* R package version 1.0-3.
URL: *http://cran.r-project.org/src/contrib/Descriptions/vcd.html* 98

Miaou, S. (1994), 'The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions', *Accid. Anal. and Prev.* **26**, 471–482. 2, 13

Miaou, S.-P. and Lum, H. (1993), 'Modeling vehicle accidents and highway geometric design relationships', *Accid. Anal. and Prev.* **25**, 689–709. 18

Miaou, S.-P. and Song, J. (2005), 'Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence', *Accid. Anal. and Prev.* **37**, 699–720. 3, 13, 168, 169

Miaou, S., Song, J. J. and Mallick, B. K. (2003), 'Roadway traffic crash mapping: A space-time modelling approach', *Journal of Transportation and Statistics* **6**(1), 33–57. **URL:** *http://www.bts.gov/publications* 3, 6, 27, 31, 39, 59

Milton, J. and Mannering, F. (1998), 'The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies', *Transportation* **25**, 395–413. 2, 13, 21

Mollié, A. (1996), Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman and Hall, London, pp. 359–379. 43, 46

Mountain, L., Fawaz, B. and Jarrett, D. (1996), 'Accident prediction models for roads with minor junctions', *Accid. Anal. and Prev.* **28**, 695–707. 3, 13

Mountain, L., Jarrett, D. and Fawaz, B. (1995a), 'The safety effects of highway engineering schemes', *Proc. Institution of Civil Engieers: Transport* **111**, 298–309. 3

Noland, R. and Quddus, M. (2004), 'A spatially disaggregate analysis of road casualties in england', *Accid. Anal. and Prev.* **36**, 973–984. 4, 76, 79

Office for National Statistics (2001), *The census in England and Wales*, The Office for National Statistics.
URL: *http://www.statistics.gov.uk/census/* 83

Office for National Statistics (2007), *Population statistics*, National Statistics.
URL: *http://www.statistics.gov.uk/* 161

Ogden, K. W. (1996), *Safer Roads: A Guide to Road Safety Engineering*, Avebury Technical. 2, 12

Okabe, A. and Yamada, I. (2001), 'The k-function method on a network and its computational implementation', **33**, 271–290. 28

Ordnance Survey (2007), *the Ordnance Survey*, Ordnance Survey.
URL: *http://www.ordnancesurvey.co.uk* 84

Pickering, D., Hall, R. D. and Grimmer, M. (1986), Accidents at rural t-junctions, Technical Report RR65, Transport Research Laboratory. 16, 18

Qin, X., Ivan, J. N. and Ravishanker, N. (2003), 'Selecting exposure measures in crash rate prediction for two-lane highway segments', *Accid. Anal. and Prev.* **36**, 183–191. 3

Richardson, S., Best, N. and Abellan, J. J. (2006), 'Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (uk)', *Stat Methods Med Res* **15**, 385–407. 51, 67, 125

Saccomanno, F. and Buyco, C. (1988), 'Generalized loglinear models of truck accident rates', *Transport. Res. Record* **1172**, 23–31. 18

Schabenberger, O. and Gotway, C. (2005), *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC. 26, 28, 182

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464. 37

Secretary of State for the Environment (1994), *The Local Government Changes for England Regulations*, The Secretary of State for the Environment.
URL: *http://www.opsi.gov.uk/SI/si1994/Uksi_19940867_en_2.htm* 82

Silverman, B. (1997), *Density Estimation for Statistics and Data Analysis*, CRC Press. 28

Song, J., Ghosh, M., Miaou, S. and Mallick, B. (2006), 'Bayesian multivariate spatial models for roadway traffic crash mapping', *Journal of Multivariate Analysis* **97**, 246–273. 5, 31

SPECTRUM (2007), *SPECTRUM*, Mott MacDonald Ltd.
URL: *http://www.mottmac.com* 84

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* **64**, 583–639. 37

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003), *WinBUGS Version User Manual*, MRC Biostatistics Unit. version 1.4.
URL: *http://www.mrc-bsu.cam.ac.uk/bugs* 35, 38, 60, 70

Sturtz, S., Ligges, U. and Gelman, A. (2005), 'R2winbugs: A package for running WinBUGS from R', *Journal of Statistical Software* **12**(3), 1–16.
URL: *http://www.jstatsoft.org* 35

Summersgill, I., Kennedy, J. and Baynes, D. (1996), Accidents at three-arm priority junctions on urban single-carriageway roads, Technical Report TRL184, Transport Research Laboratory. 3, 16

Summersgill, I., Kennedy, J., Hall, R., Hickford, A. and Barnard, S. (2001), Accidents at junctions on one-way urban roads, Technical Report TRL510, Transport Research Laboratory. 3

Summersgill, I. and Layfield, R. E. (1996), Non-junction accidents on urban single-carriageway roads, Technical Report TRL183, Transport Research Laboratory. 16

Sun, D. C., Tsutakawa, R. K. and Speckman, P. L. (1999), 'Posterior distribution of hierarchical models using car(1) distributions', *Biometrika* **86**, 341–350. 45

Taylor, M., Baruya, A. and Kennedy, J. (2002), The relationship between speed and accidents on rural single carriageway roads, Technical Report TRL551, Transport Research Laboratory. 13

Taylor, M., Hall, R. and Chatterjee, K. (1996), Accidents at 3-arm traffic signals on urban single carriageway roads, Technical Report TRL135, Transport Research Laboratory. 16

Thomas, A., Best, N., Arnold, R. and Spiegelhalter, D. (2004), *GeoBUGS User Manual*, MRC Biostatistics Unit. version 1.2.
URL: *http://www.mrc-bsu.cam.ac.uk/bugs* 60

Tunaru, R. (2001), 'Models of association versus causal models for contingency tables', *Journal of the Royal Statistical Society: Series D (The Statistician)* **50**, 257–269. 13

Tunaru, R. (2002), 'Hierarchical Bayesian models for multiple count data', *Austrian Journal of Statistics* **31**, 221–229. 5, 6, 27, 39, 49, 168, 169

Upton, G. and Fingleton, B. (1985), Spatial autocorrelation, *in* 'Spatial Data Analysis by Example', John Wiley Sons. 29, 52, 61, 71

Vidal Rodeiro, C. and Lawson, A. (2002), An analysis of edge effects in disease mapping, *in* 'GeoHealth 2002', Victoria University of Wellington. 53, 54

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997), 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**(438), 607–617.
URL: *citeseer.ist.psu.edu/waller96hierarchical.html* 50

Walmsley, D., Summersgill, I. and Binch, C. (1998), Accidents on modern rural single carriageway trunk roads, Technical Report TRL336, Transport Research Laboratory. 16

Walmsley, D., Summersgill, I. and Payne, A. (1998), Accidents on modern rural dual

carriageway trunk roads, Technical Report TRL335, Transport Research Laboratory. 3, 16

Wedderburn, R. (1974), 'Quasi-likelihood functions, generalized linear models and the gauss-newton method', *Biometrika* **61**, 439–447. 19

Wright, C. C., Abbess, C. and Jarrett, D. (1988), 'Estimating the regression-to-mean effect associated with road accident black spot treatment: Towards a more realistic approach', *Accid. Anal. and Prev.* **20**, 199–214. 13, 23

Yamada, I. and Thill, J. C. (2002), An empirical comparison of planar and network k-function analyses, *in* 'AAG Annual Meeting', Los Angeles, CA. 28

Zeileis, A. and Hornik, K. (2006), *HCL-based Color Palettes in R*. 98