University of Southampton

Faculty of Social and Human Sciences

School of Mathematics

Modelling Trends in Road Accident Frequency – Bayesian Inference for Rates with Uncertain Exposure

by

Louise Kate Lloyd

Thesis for the degree of Doctor of Philosophy

March 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

SCHOOL OF MATHEMATICS

Doctor of Philosophy

MODELLING TRENDS IN ROAD ACCIDENT FREQUENCY – BAYESIAN INFERENCE FOR RATES WITH UNCERTAIN EXPOSURE

Louise Kate Lloyd

Several thousand people die as a result of a road accident each year in Great Britain and the trend in the number of fatal accidents is monitored closely to understand increases and reductions in the number of deaths. Results from analysis of these data directly influence Government road safety policy and ensure the introduction of effective safety interventions across the country.

Overall accident numbers are important, but when disaggregating into various characteristics, accident risk (defined as the number of accidents relative to an exposure measure) is a better comparator. The exposure measure used most commonly for accident rate analysis is traffic flow which can be disaggregated into vehicle types, road type, and year. Here we want to assess the accident risk across different car types and car ages, and therefore alternative exposure sources are required. We disaggregate exposure to a further extent than possible with currently available data in order to take the increased variability within these new factors into account.

Exposure data sources are mainly based on sample surveys and therefore have some associated uncertainty, however previous accident risk analysis has not, in general, taken this into account. For an explicit way to include this uncertainty we use a Bayesian analysis to combine three sources of exposure using a log-Normal model with model priors representing our uncertainty in each data source.

Using further Bayesian models, we propagate this uncertainty through to accident rates and accident severity, determining important factors and interrelationships between factors to identify key features affecting accident trends, and we make the first exploration of the effect of the recent recession on road accidents.

Contents

List of Tables

List of Figures

Author's Declaration

I, Louise Kate Lloyd, declare that the thesis entitled

Modelling Trends in Road Accident Frequency – Bayesian Inference for Rates with Uncertain Exposure,

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

- part of this work has been published as Lloyd et al. (2013);
- part of this work is being considered for publication in the Computational Statistics and Data Analysis journal as Lloyd and Forster (in press).

Signed:..... Date:.....

Acknowledgements

I would like to thank my supervisor Professor Jon Forster for his help with completing this thesis. I also acknowledge the Transport Research Laboratory (TRL) for funding the research, the Department for Transport for the use of the On The Spot database, STATS19 and traffic data, the DVLA for providing registered vehicle data and Maureen Keigan and Jeremy Broughton at TRL for helpful advice.

Chapter 1

Introduction

1.1 Application

In recent years, around 2 000 people have died each year as the result of a road accident in Great Britain. The trends in numbers of deaths and seriously injured casualties are monitored closely by the Department for Transport, and Government road safety policy is guided by these trends.

In the late 1990s there appeared to be a change in a long running trend: the trends in fatal and serious casualties, which had been similar until that point, began to diverge, with the number of serious casualties reducing at a faster rate than the number of fatalities. In particular the number of car occupant deaths remained fairly constant for some years at around 1 700 each year (see Figure 1.1). As a consequence, the severity rate (proportion of fatalities in killed and seriously injured: KSI) of car occupants rose gradually over this period to a peak of 11% by 2006. In 2007 a rapid decrease in the number of fatally injured casualties began with numbers dropping over four consecutive years.

It was observed (Broughton and Buckle 2007) that from the mid 1990s to 2006,

several types of accidents showed an increase and were therefore particularly influential during the 'stationary' period. These accident types included:

- Single vehicle accidents: the proportion of fatally or seriously injured car occupants involved in single vehicle accidents increased and the number of fatally injured car occupants in single vehicle accidents increased;
- Accidents at bends: the proportion of fatally or seriously injured car occupants involved in accidents at bends increased;
- Accidents where a car left the carriageway: the proportion of casualties whose cars left the carriageway increased (a slight decrease in 2006);
- Accidents where a car overturned: the proportion of fatally or seriously injured car occupants injured when their car overturned increased;
- Accidents involving a young driver: the number of young drivers killed increased;
- Accidents involving a large car: an increase in fatality rate (relative to car fleet) of 4x4 and people carrier occupants (and small saloons) was observed;
- Accidents involving an old car: a general increasing trend (which has stabilised in the last few years) of car occupant fatalities in old cars was observed.

In a comprehensive univariate analysis presented in Broughton and Buckle (2007) all the variables (single- or multi-vehicle accident, accident occurred at bend, type of car involved, age of driver, etc.) that describe the accident types listed above appear to be affecting the fatality trend. These variables are known to be correlated, for example it is likely that vehicles are more likely to overturn if they leave the road, and younger, less experienced drivers are more likely to drive older cars. In assessing these variables individually, important inter-relationships between the variables are likely to be missing. These separate analyses need to be combined in order to determine which of them are driving the trend, taking into account correlations between the variables.

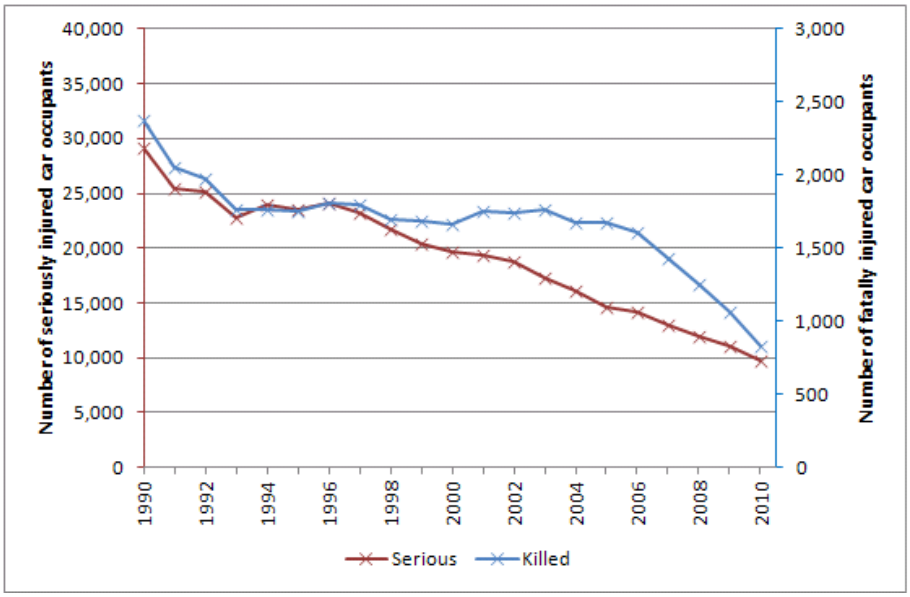


Figure 1.1: Annual numbers of fatally (blue) and seriously (red) injured car occupants from 1990 – 2010

Since 2006 the trend in fatal accidents has changed again, dropping quickly and consistently. The reasons for this are unclear and there has been limited research to investigate; one suggestion is that the drop is related to the economic recession (see section 1.3.5). More up to date analysis is required to see if these accident types have been particularly influential in this recent downward trend in fatal accidents. We use a combination of accident and exposure[1] data to assess all these factors together.

1.2 Accident data

STATS19 is the system for recording personal injury road accident data reported to the police in Great Britain. The database is hierarchical, comprising details of the accident circumstances, within that, data on each vehicle involved and then information collected at the casualty level. The data are collected by Police

[1]Exposure is a measure of the number of times road users are exposed to a potential accident

Officers and also include the officer's subjective view of the contributing factors, or possible reasons why the accident occurred.

Around 70 variables make up the STATS19 record, with some repeats for multi-vehicle and multi-casualty accidents. The research detailed in this thesis is based on a subset of the STATS19 data containing killed or seriously injured car occupant casualties involved in single vehicle accidents from 1999–2010. This subset contains information on 45 394 accidents.

The applicable variables used in the research are detailed below.

Year: the year the accident took place. Data are from 1999 to 2010.

Severity: the highest severity of the car occupant casualties involved in the accident. Data are killed or seriously injured.

Overturn: binary variable to denote cars that did and did not overturn during or as a result of the accident.

Bend: binary variable to denote whether the accident occurred at a bend.

Road type: type of road on which the accident occurred. Data are Motorway, A road or Minor road.

Car type: the type of car involved in the accident is split into six groups by size of vehicle. This information is retrieved from the number plate data recorded in STATS19 which is linked to a DVLA database and is complete in around 80% of accidents, with varying completeness per year.[2] Data are minis and superminis, small saloon, medium saloon, large saloon, 4x4s and people carriers, and sports cars.

Car age: the age of the car is derived from the registration number in the accident data and DVLA database. The age groups used here are 0–2 years, 3–5 years, 6–10 years, 11–15 years and 16 or more years.

[2]Cases where the car type is not known are randomly allocated a car type in order to model the overall accident rate rather than a restricted unbalanced dataset.

The STATS19 database is a very valuable resource and is used extensively in research aimed at reducing the number and severity of road accidents. However it is, and has to be treated as, only part of the story. Analysis of STATS19 in isolation gives a clear picture of the number, severity and location of road accidents in Great Britain. It does not provide any information on how accident numbers relate to general (non accident involved) traffic patterns or other external factors. Exposure data are required in order to put accident data into context.

1.3 Exposure data

Cullen and Frey (1999) describe exposure data as the combination of information about the frequency, intensity and duration of contact with risk. In particular, exposure to road accidents is the term used to describe a measure of the potential that a group of subjects have to be involved in an accident.

The most common type of exposure data for road accident analysis is the number of vehicle-kilometres travelled each year. One vehicle-kilometre is defined as one vehicle travelling one kilometre. Data such as number of registered vehicles, length of road network and total fuel consumption for road transport (Van den Bossche and Wets 2003) are also used. Its main use and context is to estimate risk of being involved in an accident, and it is risk that often helps to prioritise safety measures. For example, the number of motor vehicle occupants/riders killed or seriously injured (KSI) in 2010 was 16 134 (Department for Transport 2011) of which a majority (60%) were car occupants. However, when the numbers of kilometres driven on the different road types are taken into account (and assuming that each vehicle type has a similar distribution of occupants), the casualty rate (relative to the number of vehicle kilometres driven) for car occupants is considerably smaller (25 KSI car occupant casualties per billion car km) than the casualty rate for other vehicles (62 KSI other vehicle occupant casualties per billion other

vehicle km). So, even though there is a higher number of car occupants killed or seriously injured than other vehicle occupants, they travel many more kilometres per year and therefore the risk of being seriously injured or killed in an accident is considerably higher for occupants of other vehicles than cars. The necessity for exposure data when assessing road accident numbers is further demonstrated in Section 1.4.

Chapter 2 contains a further discussion on exposure data, and the combination of different sources of exposure data to achieve a better estimate of overall exposure is reported in Chapter 4. Various sources are described in Sections 1.3.1 to 1.3.4.

1.3.1 Traffic data

The Department for Transport collects and analyses traffic count data on a large selection of roads in Great Britain. These counts are combined with road network lengths in order to estimate the total vehicle kilometres travelled each year. Overall, there are approximately 50 000km of Motorway and A roads and 344 000km of Minor roads (Department for Transport 2011). Traffic densities vary considerably over different types of roads, different areas and different vehicle types, so detailed data are collected automatically and manually at a large number of sites across Great Britain. Traffic flow, measured in vehicle kilometres, is the product of the average daily flow (measured in vehicles and calculated from the traffic count) and the length of the road on which the daily flow was based. Due to the nature of the counting mechanisms, traffic flow can be approximately disaggregated by time, month, road type, region and vehicle type, however a disaggregation by different car types is not available.

The data used in this research are car traffic flow from 1999–2010 on different road types, shown in Table 1.1.

Table 1.1: Annual car traffic flow in Great Britain (billion vehicle kilometres) by road type from 1999 – 2010

Year	Motorway	A roads	Minor roads
1999	63	174	143
2000	71	165	143
2001	72	168	144
2002	70	177	145
2003	70	179	144
2004	73	181	145
2005	73	180	145
2006	74	181	142
2007	75	178	145
2008	75	177	143
2009	75	178	141
2010	74	175	137

1.3.2 Registered vehicle data

Car type

The Driver & Vehicle Licensing Agency (DVLA) holds information on each registered vehicle in the UK, including the make and model of the vehicle. The data that have been used in this research are the number of registered vehicles by car type and year from 1999–2010 (Table 1.2). This make and model information can be used to categorise the UK car fleet into six subgroups by car body type: minis and superminis, small, medium and large saloons, 4x4s and people carriers and sports cars. Approximately 76% of all the known make and model combinations in the registered cars dataset have been classified into a car type subgroup and this covers 99.5% of all registered cars.

Car age

The DVLA database also holds information on the year of registration of each vehicle, and from this we can derive the age of the cars registered. The age of a

Table 1.2: Number of registered cars in Great Britain by type and year (millions)

Year	Minis & superminis	Small saloons	Medium saloons	Large saloons	4x4 & people carriers	Sports cars
1999	6.5	7.8	5.8	1.9	1.0	0.6
2000	6.7	7.8	5.8	1.8	1.1	0.6
2001	7.0	7.9	5.9	1.8	1.3	0.7
2002	7.3	8.0	5.9	1.7	1.6	0.8
2003	7.6	7.9	5.9	1.7	1.8	0.9
2004	7.9	7.9	5.8	1.7	2.1	0.9
2005	8.1	7.9	5.7	1.7	2.4	1.0
2006	8.2	7.9	5.5	1.7	2.6	1.0
2007	8.4	7.9	5.4	1.8	2.9	1.0
2008	8.6	7.9	5.2	1.8	3.0	1.0
2009	8.7	7.8	4.9	1.8	3.2	1.0
2010	8.9	7.8	4.7	1.8	3.3	1.0

vehicle often affects the severity or likelihood of an accident due to the continuous development in primary and secondary safety[3] features installed in vehicles (Broughton 2003). Over a period of years new technologies and designs are introduced which means that new cars in 2010 generally have more safety features than new cars in 1999. Broughton (2003) shows that the design year of a car affects the proportion of occupants who are killed or seriously injured (KSI) in collisions, with more modern cars having a lower proportion of KSI occupants than older cars. We have used vehicle age as the key variable to represent safety feature design improvements and we can derive design year from the interaction between vehicle age and year. Figure 1.2 shows the changing distribution of registered car ages within the different car types over the 12 year period of interest.

[3]Primary safety refers to (mainly recent) technologies designed to automatically avoid a crash such as electronic stability control and automatic braking systems. Secondary safety in vehicles includes technologies, such as seatbelts and airbags, and structural design intended to reduce the impact of a collision once it has occurred.

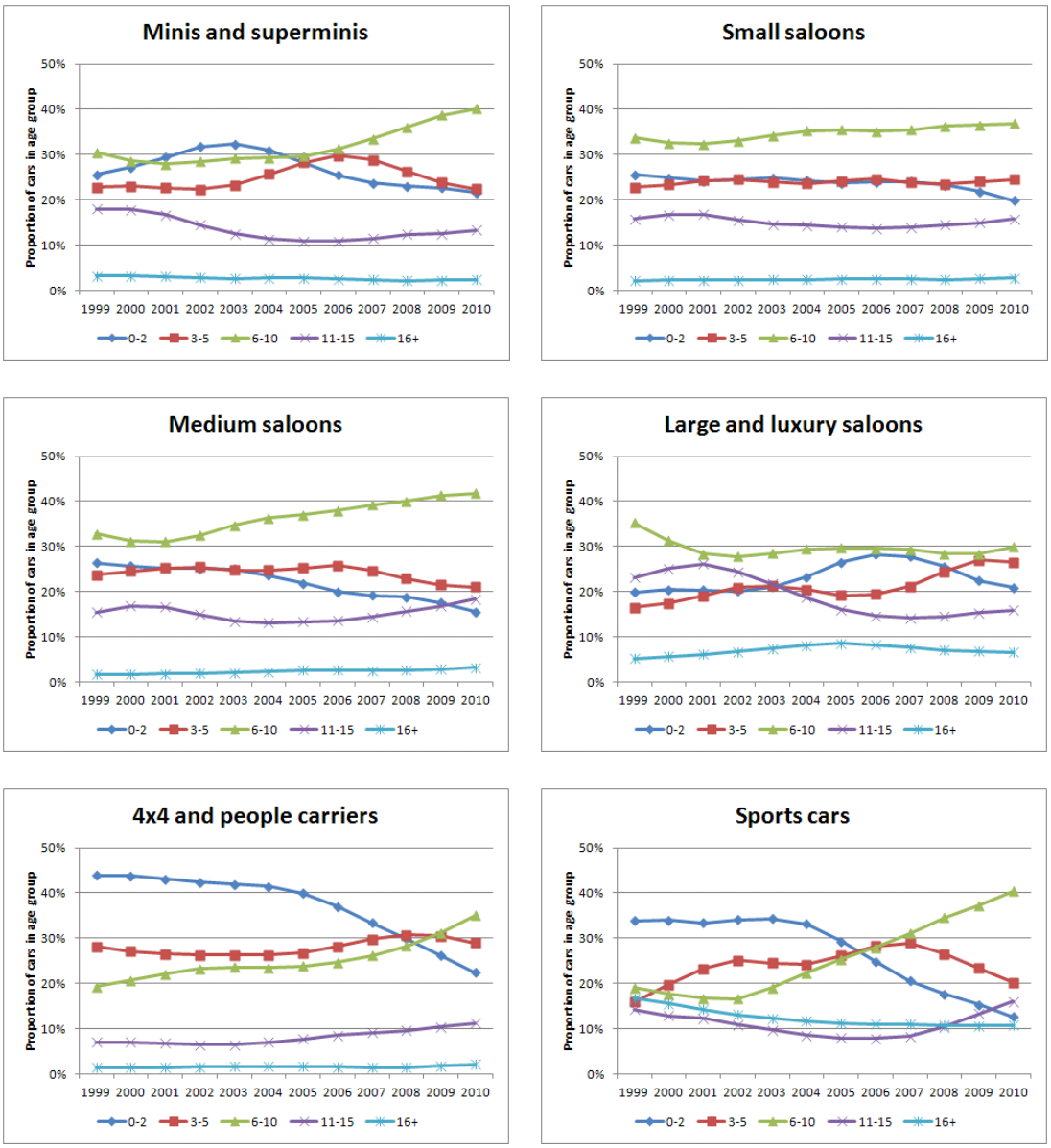


Figure 1.2: Distribution of car ages by year for each car type from 1999 – 2010

1.3.3 Induced exposure data

Classical exposure measures such as vehicle kilometres or number of registered vehicles are generally used to calculate accident rates. These sources of data are restrictive as only limited disaggregation is possible and they do not represent well the variation in different driver populations in different types of vehicles on different road types.

af Wahlburg and Dorn (2007) summarise the reasons for not using mileage as exposure data:

- high mileage drivers have a lower crash risk per mile than low mileage drivers as
  - high mileage drivers mostly drive on safe highways
  - low mileage drivers drive on busy two-way streets
  - high mileage drivers possess better driving and safety skills as they are more experienced
- disaggregating mileage as an exposure measure is difficult for most factors as different types of drivers drive on different roads and at different times, in different aged/types of vehicles.

Thorpe (1967) introduced induced exposure techniques to deal with some of these problems. He defined the relative likelihood of driver involvement in accidents as the ratio of involvement to exposure. Exposure was calculated by $2M_i - S_i$, where S_i is the percentage of single vehicle accidents and M_i is the percentage of multi-vehicle accidents for a specific group of drivers i. Carr (1970) adapted the concept by classifying a driver as responsible, i.e. 'at-fault' or 'not-at-fault' in a multi-vehicle accident based on the police report. Not at fault is commonly defined as 'without any contributing human factors to the crash occurrence as defined by the police investigator officer or accident investigator' (Chandraratna

and Stamatiadis 2009). Exposure as defined here is the percentage of not at fault drivers in multi-vehicle accident for group i and it is assumed that these drivers are a random representation of drivers on the road at the time of the accident.

This is a powerful alternative for certain exposure measures which are not available directly, in particular for calculating accident risk of drivers with particular characteristics. However, researchers are reluctant to use induced exposure techniques as a full comparison with conventional methods has not been carried out, and the underlying assumptions are not always justified. A review of the assumptions is given in Appendix A.

Several papers have used the induced exposure technique to evaluate accident propensity.

Redondo-Calderon et al. (2001) used Carr's (1970) classical quasi-induced exposure method to compare risk among different driver categories under different types of environmental conditions. Spanish road accident data were used from the two-year period 1991–1992 and results show that crash risk was 1.4–2.4 times greater in men than women, significantly higher in younger drivers and for those under abnormal psychophysical conditions.

Difficulties arose with small numbers – some categories had to be combined and only a few variables could be used at any one time, however, it was still deemed an easy and economical tool for estimating accident risk.

Yannis et al. (2005) have also used this technique to provide exposure data in order to assess the accident rates of young motorcyclists, and to assess the relative accident risk of foreign drivers in Greece (Yannis et al. 2007).

1.3.4 OTS data

On-The-Spot (OTS) accident research (Cuerden et al. 2008) was an in depth accident investigation and data collection project which ran from 2000 to 2010. Two teams of trained accident investigators based in Berkshire and Nottinghamshire attended a sample of road accidents, occurring in their local area, to collect detailed data including the circumstances of the accident, the weather, light and road conditions, vehicle damage and vehicle occupant information; this is information which often disappears very quickly after the accident. Additional information was collected from hospitals and coroners and through questionnaires sent to crash participants. In total data from around 4 500 accidents were collected.

The OTS data used in this research have been provided by the Department for Transport and are an example of induced exposure data. We assume that the drivers deemed to be not at fault in all OTS multi-car accidents were random representatives of the drivers and associated car types on the road at the time of the accidents. The distribution of car type by road type is shown in Table 1.3. As the numbers are quite small, disaggregation by year was not possible. The study sampling procedure prioritised high speed collisions, and therefore it is not representative sample of collisions on different road types. We use the distribution of car types within each road type as shown in the second half of Table 1.3.

1.3.5 Economic data

It has been suggested (Broughton 2009) that the economic position of the country has an effect on the number of road accidents that occur. This could be for a number of reasons, including drivers tending to choose to drive less or at more economical speeds in times of recession. There is little research into the link between economy and road accidents but it does appear to explain unusual drops in road accidents in the early 1980s, 1990s and 2008–2009 whilst the UK was in

Table 1.3: Number of not-at-fault car drivers by car type and road type in multi-car accidents in the OTS database

Car type	Motorways	A roads	Minor roads
Minis & superminis	52	159	68
Small saloons	72	143	80
Medium saloons	61	116	52
Large saloons	15	43	8
4x4s & people carriers	21	51	23
Sports cars	10	18	10
Minis & superminis	23%	30%	28%
Small saloons	31%	27%	33%
Medium saloons	26%	22%	22%
Large saloons	6%	8%	3%
4x4s & people carriers	9%	10%	10%
Sports cars	4%	3%	4%

recession. We will investigate this link in Chapters 5 and 6 using Gross Domestic Product (GDP) as an explanatory variable. GDP is a measure of the UK's economic activity. The data used in this research is nominal GDP per capita which is the average value of production per person per year at current prices. These data have been retrieved from MeasuringWorth (Officer and Williamson 2010) and are shown in Table 1.4.

1.4 Accident and exposure data trends

Section 1.1 refers to the accident types which were particularly influential in the stationary fatal trend from the mid 1990s to 2006 and this includes an increase in the number of accidents involving large cars such as 4x4s and people carriers. The obvious explanation for this is that there was an increase in the number of 4x4s and people carriers being bought and therefore driven on the roads over the same period. How much of the increase in accidents that can be explained by the increase in exposure is to be shown later, but this section gives a short summary of

Table 1.4: Nominal GDP per capita by year from 1999 – 2010

Year	Nominal GDP per capita (£k)
1999	15.8
2000	16.6
2001	17.3
2002	18.1
2003	19.1
2004	20.1
2005	20.8
2006	21.9
2007	23.1
2008	23.4
2009	22.6
2010	23.5

the data and develops the idea of a need for a viable exposure data to disaggregate by car type.

Figure 1.3 shows the differing distribution of type of car involved in single vehicle accidents where an occupant is killed or seriously injured by year. In general, from a 1999 baseline, there is a proportional increase in accidents involving 4x4s, minis and sports cars and a decrease in the proportion of accidents involving large, medium and small saloons over the 12 year period under investigation.

Without including exposure data in the model it is impossible to determine if this effect is due to accident distributions changing or just due to a change in vehicles on the road network (i.e. exposure). The distribution of different car types registered each year (described in Section 1.3.2) is one type of exposure which disaggregates car types. Figure 1.4 shows that this distribution appears to have changed in a similar way to that seen in Figure 1.3 for the accident data, i.e. the proportion of registered cars that are minis, sports cars and 4x4s has increased over the 12 years, and the proportion of the other cars types has decreased. In fact, there are three times the proportion of 4x4s registered in 2010 as there were

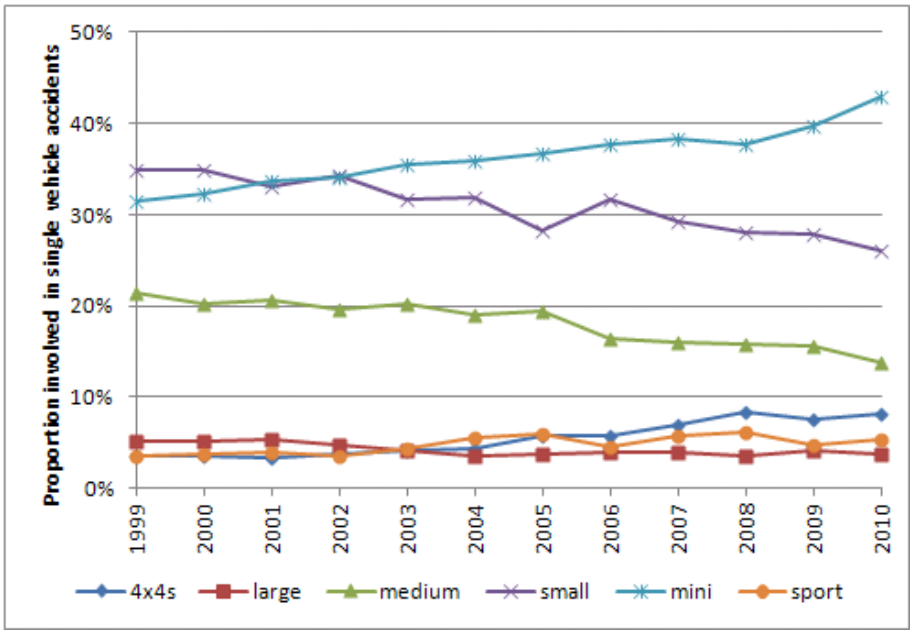


Figure 1.3: Distribution of single vehicle accidents by car type for each year

in 1999.

It is clear from Figures 1.3 and 1.4 that the changing patterns of accident trends and one type of exposure data are the same. However, we need to check if these trends are changing at a similar rate.

Combining these sources of data together, Figure 1.5 shows the 'proportional accident rate' per year. Each line represents the proportion of accidents in a particular car type divided by the proportion of that car type registered in Britain. Values higher than one represent a higher proportion of accidents than the proportion of that car type on the road. This has consistently been the case for minis, small saloons (until 2004) and sports cars. This rate has remained fairly constant for small, medium and large saloon cars and 4x4s and risen for minis and sports cars.

This steady trend for 4x4 cars may suggest that the rise in the proportion of 4x4s in the car fleet accounts for the rise in proportion of accidents involving 4x4s in the same period. For minis and sports cars it is a different story: when comparing

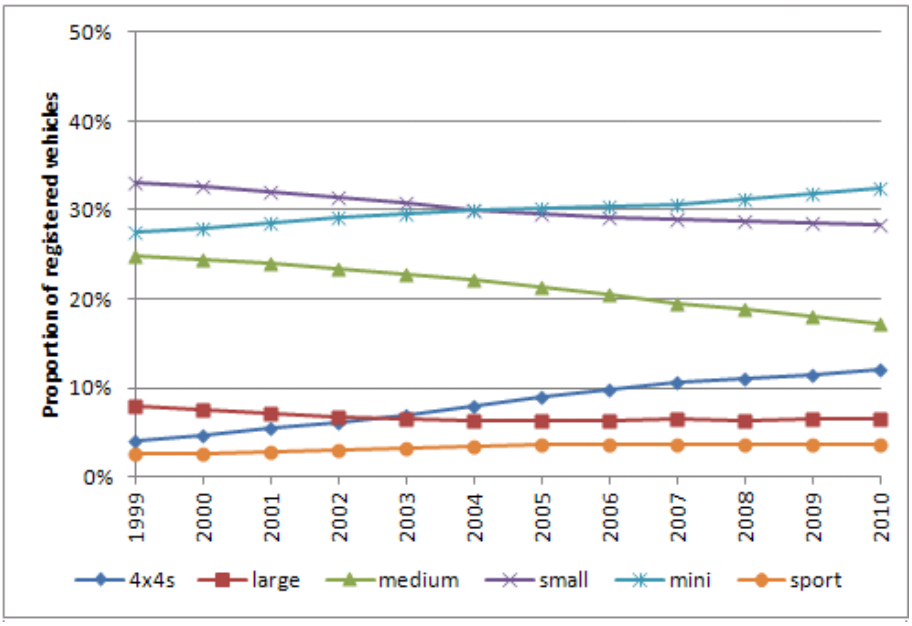


Figure 1.4: Distribution of car types registered in Great Britain each year

increases in the proportion of accidents involving minis and sports cars in Figure 1.3 with smaller relative increases in the proportion of the fleet that are minis and sports cars in Figure 1.4, an increasing rate is observed. This implies that, based on registered vehicle numbers as an exposure measure, each mini and sports car is getting proportionally more likely to be involved in an accident.

These patterns are interesting, but registered vehicle data are not an ideal exposure measure as a single source: different car types will tend to be exposed to different risks, for example, particular car types may be used more often for longer journeys. The observed patterns in Figure 1.5 do not take into account relative use of the road network by different car drivers, and therefore the picture is not yet complete.

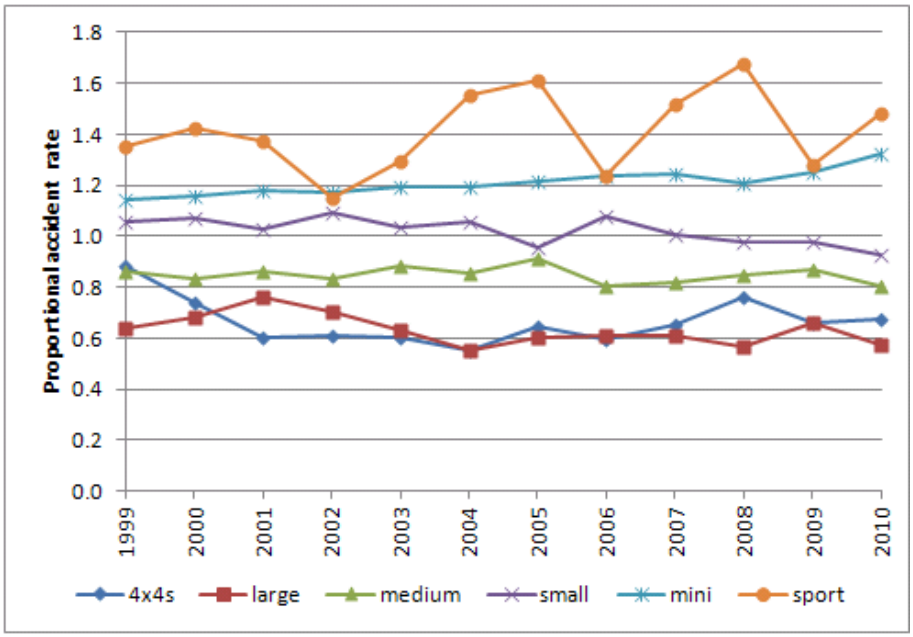


Figure 1.5: Proportional accident rate per proportion of car types registered each year

1.5 Aim

The overall aim of this research is to evaluate the changes in the fatal and serious road accident trends by modelling accident and exposure data simultaneously.

This will require development of a novel statistical methodology to account for uncertainty in exposure data and combining datasets to achieve better estimates of the effect of influential factors.

The analysis will be based on a subset of accidents seen to be affecting the trend: that is, those involving only one car where an occupant was killed or seriously injured. A combination of accident, exposure and explanatory data will be used to achieve this aim.

1.6 Thesis structure

There are three main parts to this thesis. Part 1 contains Chapters 2, 3 and 4 and discusses modelling exposure data. Chapters 2 and 3 describe the limitations of exposure data and a review of Bayesian methods respectively. Chapter 4 uses Bayesian methods to combine the sources of exposure data described in Sections 1.3.1 to 1.3.4 to answer some of the limitations described in Chapter 2.

Part 2 (containing Chapters 5 and 6) takes the modelled exposure data from Part 1 and builds models for accident rates and accident severity. The accident rate model predicts the chance of having a KSI accident in different car types, on different road types and in different years. Once an accident has occurred, the accident severity model predicts how severely the car occupants were likely to be injured given their car type and car age, where the accident occurred and whether the car overturned. An assessment of the effect of the recession is also made in these chapters.

Part 3 (Chapter 7) discusses Graphical Modelling techniques and uses these techniques to draw all three models together and predict future trends.

Chapter 8 summarises the research.

Chapter 2

Unknown and Variable Exposure

2.1 Introduction

Every time someone uses the road network (either in or on a vehicle, or as a pedestrian when there is a vehicle in close proximity) they are exposed to the risk of being involved in an accident. Individual accident risk is related to your own exposure to accidents – how often, when and how you travel on the road network. As road accidents are relatively rare events and due to the fact that measuring each individual's exposure is an impossible task we consider the accident risk of groups of people.

Accident risk is defined as

$$\text{accident risk} = \frac{\text{accident propensity}}{\text{exposure}}$$

where propensity is the raw number of accidents. Accident risk and propensity are used together to understand road safety priorities and to inform government policy for reducing road casualties. A high propensity of accidents in a particular group

of road users may purely reflect a high number of people within that particular road user group. A large demographic group, for example females aged under 80 in Berkshire will be involved in a larger number of accidents than a small demographic group such as females aged over 80 in Berkshire, due, at least in part, to the number exposed to accidents in the two groups. Therefore risk (and consequently exposure) is an essential consideration.

Figure 2.1 demonstrates a possible categorisation of the accident risk and propensity for different road user groups. Of course these quantities are measured on a continuous scale, however this demonstrates the concept.

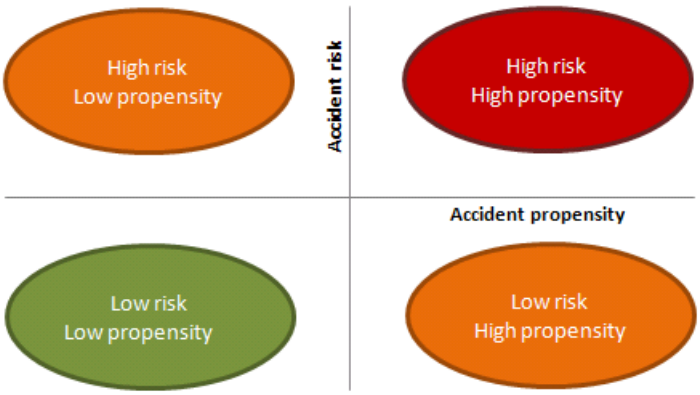


Figure 2.1: Idealistic relationship between accident rate and accident propensity

Road user groups with a high accident risk and a high propensity of accident involvement are the first priority in road safety strategy. These are large groups which are not only involved in a high proportion of the country's accidents, but also, once the size of the groups and their road use exposure is taken into account, individuals in these groups are more likely, on average, to be involved in an accident than the average road user – they may be less risk averse, or perhaps less experienced. Groups which fall into this category can be targeted by a mixture of education, enforcement and engineering strategies to reduce their risk. Just reducing the risk by a small amount for this group has a large effect on the number

of accidents.

High risk and low propensity groups and high propensity and low risk groups are considered next in government policy. Reducing the involvement in accidents of these two categories requires different treatment. It is relatively easy to identify interventions which could reduce the risk of individuals in a high risk group (although it may not be financially attractive), and it is also not complicated to disseminate information or to introduce interventions to affect large groups of people. However, affecting a small number of high risk individuals (high risk, low propensity) or identifying interventions for a large group of low risk individuals (low risk, high propensity) is very difficult.

Road user groups classified in the green segment (low risk and low propensity) are often left – they contribute just a small number of accidents to the overall total, and individuals in the group are, on average, less likely to have an accident than the average road user, once their exposure has been taken into account.

If we cannot measure individual risk then the biggest number of different road user groups possible must be the starting point for determining the position of these groups within the rate propensity graph shown in Figure 2.1. As described in Section 1.2, the limitation for disaggregating road user types is not accident data – detailed information is collected at each accident attended by the police. The limitation is in the exposure data.

Exposure is a difficult concept to measure. Commonly, a measure of the number of vehicle kilometres (discussed in Section 1.3.1) is used as a proxy. This measure can be disaggregated by road type, year and vehicle type as described in Department for Transport (2010). In order to evaluate the hypotheses of Broughton and Buckle (2007) discussed in Section 1.1, however, further disaggregation of this exposure proxy is required, in particular by:

- age of driver;
- type of car; and
- age of car.

Information on these variables can be sought in different datasets, for example the registered vehicle dataset described in Section 1.3.2 includes information on the number of cars of different body types and age, the OTS induced exposure data (Section 1.3.4) suggests the distribution of different cars using different road types and the National Travel Survey (Department for Transport 2012) contains data on the age and sex of car drivers. Up until now accident rate modelling research has been based on one source of exposure data, for example: Knowles et al. (2007) use the number of registered cars by car type to evaluate whether people carriers and 4x4s are involved in relatively more accidents than other types of car; Broughton and Knowles (2010) use the National Traffic estimates in vehicle kilometres to evaluate the overall progress towards the Government's road safety target for 2010; Sonkin et al. (2006) use estimates of the number of miles travelled by children by mode of transport from National Travel Surveys to evaluate the deaths amongst children walking and cycling; and Fridstrom and Ingebrigtsen (1991) use fuel sales to evaluate the contribution to accident risk made by a series of explanatory factors such as weather, economy, seatbelt use and law enforcement.

A combination of some of these data will be attempted in order to derive an approximate disaggregated exposure dataset useful for evaluating the trend in the effect on accident risk for a much bigger set of road user groups than is possible using just one of these datasets. This is discussed further in Section 2.2.1 below.

An additional consideration is the uncertainty about these data – most of these datasets are based on sample surveys: vehicle kilometre data are based on a mixture of automatic and manual surveys across a sample of major and minor roads and are combined to form the National Traffic estimates (Department for Trans-

port 2010); the National Travel Survey combines information from interviews and travel diaries for around 8 000 people each year to produce estimates for the whole of Britain; and the OTS data are collected at a sample of accidents attended by accident investigators – these data are believed to be the least robust. The registered vehicle data is an up-to-date census of registered cars, however some uncertainty still arises due to missing information on unregistered cars and incomplete information on scrapped cars.

The majority of research into accident rates does nothing to account for uncertainty in the exposure: in Reported Road Casualties Great Britain (Department for Transport 2011), the British compilation of road accident statistics, the traffic estimates described above are used as the exposure measure to calculate the number of accidents per billion vehicle kilometre. Modelling of accident rates is mostly based on the classical approach and believes the chosen measure of exposure (e.g. traffic, fuel consumption, induced exposure) to be fixed and true. Van den Bossche and Wets (2003), Starnes and Longthorne (2003), Tunaru and Jarrett (1998) and Yannis et al. (2007) along with the references mentioned above (Knowles et al. 2007; Broughton and Knowles 2010; Sonkin et al. 2006; Fridstrom and Ingebrigtsen 1991) all contain examples of the use of exposure information, based on survey data but assumed to be a fixed quantity, to model accident risk.

Some research has been conducted which does at least consider the uncertainty (and variability) in traffic exposure. Rosas-Jaimes et al. (2011) evaluate accident rates at a selection of sites in Toluca, in the State of Mexico, where the vehicular flow has been measured directly and continuously, and therefore can be used, they assert, with certainty. They assess a selection of Bayesian statistical models which consider uncertainty in the accident numbers, caused by systematic errors, inconsistencies and underreporting, but keep the vehicular flows fixed. Qin et al. (2006) use a similar method where they believe that use of a finer temporal disaggregation (exposure by hour) will advance the development of crash prediction models

for four crash types: single-vehicle, multi-vehicle same direction, multi-vehicle opposite direction, and multi-vehicle intersecting direction. The relationship between vehicle flow and crash occurrence appears to be different across these four crash types and this variability in the accident–exposure relationship is dealt with by applying separate binary regression models to each type. Once again, only a selection of sites where this detailed exposure data exists can be used within the accident modelling.

Research into traffic network modelling must be mentioned at this stage – this is an area of applied research evaluating and predicting traffic flows, primarily on main roads and major junctions, to feed in to live traffic management systems. These models forecast a highly disaggregated set of data, often using Bayesian models and therefore considering uncertainty in the estimates. With annual accident rate modelling in mind though, results from these models are restrictive: Queen and Albers (2009) discuss use of multiregression dynamic models to model real-time traffic flow data by hour across two junctions on the M25, London's orbital motorway, and sheer quantity of detailed data and processing requirements show the limitations of the techniques for our broader use.

The disaggregation of the exposure estimates requires the combination of several types of uncertain exposure and, as discussed above, there has been no research into combining sources of exposure data or modelling uncertainty in road accident exposure data. Here we review other areas of risk research where combining data and uncertainty and variability in exposure is considered, such as in the areas of toxicology, epidemiology and nuclear power technology.

Cullen and Frey (1999), amongst others, suggest that an understanding of the difference between variability and uncertainty, and the sources of such quantities, results in a better appreciation of the limits of the analysis, gaps in data and areas for future research. In order to fulfil this aim, definitions of the two quantities are

given below.

Variability

Van Belle (2008) refers to variability (or aleatory uncertainty) as 'natural variation in some quantity', and this natural variation could be differences in an input across time, space or between individuals. In the road exposure context this relates to, for example, time of day of driving, choice of vehicle or road type.

Variability is a concept which cannot be reduced but can be learnt about with additional data collection.

Uncertainty

Uncertainty (also known as epistemic uncertainty or fuzziness) is defined in Helton (1996) as the 'measure of incompleteness of one's knowledge or information about an unknown quantity whose true value could be established if a perfect measuring device existed'. Sources of uncertainty include measurement error, model assumptions and proxies to exposure such as vehicle kilometres which are in use here.

Uncertainty is caused by a lack of knowledge about a particular value, due often to data collection over too small a sample and therefore, in principle, uncertainty can be reduced by collecting more or better data.

2.2 Modelling exposure

Commonly, exposure is assumed to be fixed and known. It is then difficult to understand variability and uncertainty in exposure results, and, importantly, results of this form suggest over-confidence in the conclusions, and miss out on showing explicitly any needs for further data collection. Cullen and Frey (1999)

suggest that these may be serious limitations. An obvious choice is to consider a Bayesian structure which can represent variability and uncertainty explicitly. Bayesian methods are described in Chapter 3. As a brief summary, it is possible to represent variability and uncertainty in inputs as frequency or probability distributions rather than point estimates. These distributions can then be used to propagate uncertainty and variability through into further modelling such as for accident rates. It is also possible to separate these quantities within a Bayesian analysis, as described in Morgan and Henrion (1990), if it is important to be able to distinguish between them.

Cullen and Frey (1999) provide an excellent discussion of dealing with uncertainty and variability in exposure models, and some of the following discussion is based on their work.

2.2.1 Variability in exposure models

Variability comes in several forms:

- Temporal variability: Quite often exposure measures can be quantified across a series of different time scales, for example, it is possible to measure the quantity of traffic passing a point on a stretch of road over each minute, hour, day, month or year. There will be some temporal variability across these times – traffic tends to be lighter at night than during the day for example. Choosing the appropriate level at which to aggregate exposure measures in time, and therefore the distribution and associated parameters required to represent the variability, depends on the assessment question. In this case, the overall aim of the modelling described in Chapter 4 is to investigate for which groups of vehicles the annual accident rate has changed over the time period under investigation, and it is appropriate to ignore daily variability and aggregate over each year.

- Spatial variability: It is important to define the spatial coverage of the exposure information and evaluate whether variation across the space is relevant to the assessment question. It is suggested by Cullen and Frey (1999) that spatial variability can often be a surrogate for inter-individual variability. Considering geographical variation in accident rates is possible but has not been done here as the change in annual trend does not appear to be noticeably different across the different regions (Lloyd et al. 2013).

- Inter-individual variability: As described above, each individual has their own risk of being involved in an accident, dependent on a whole series of variables which may include time of day of travel, choice of vehicle, distance travelled and risk aversion level amongst many other factors all contributing to inter-individual variability. Aggregation over groups of individuals reduces the variability to be modelled and, once again, may be appropriate in relation to the research question. In this case, we hypothesise that the general decrease in fatal accident trend in the most recent years may be due to a decrease in the number of young drivers (one of the most risky groups) and therefore to disaggregate by age may be appropriate.

In epidemiological studies Loomis and Kromhout (2004) suggest that the variability types of interest are temporal and inter-individual. In this paper and other similar epidemiological research, these levels of variability are not dealt with within a Bayesian framework. Loomis and Kromhout (2004) suggest that a common simplifying assumption in this area is that homogeneous groups are uniformly exposed, based often on criteria such as job title or work area. For example, lab technicians in a particular department. They do recognise that there is variability within groups too and assess this within- and between- variability using ANOVA modelling. If the within variability is large relative to the between variability then individuals are often assigned an exposure level which is equal to the mean of the group, otherwise their individual exposure measure is used.

Alternatively a weighted mean of the individual exposure value and group mean may be used (Seixas and Sheppard 1996, gives an example of this). They suggest that if direct measurements are sparse or unreliable then empirical models could be used to augment the data. Preller et al. (1995), for example, uses multiple regression modelling to achieve this from a set of readily observable factors such as location, activity or job title.

There is a substantial amount of temporal, spatial and inter- and intra-individual variability within the traffic exposure data – different people use different roads at different times of day and different lengths of journey. Similarly to Loomis and Kromhout (2004) we do not deal with this variability within a Bayesian framework. The data have been disaggregated over a set of variables to account for variability within car body types and car ages. We assume that this accounts for any significant variability in exposure. Use of additional data and further data collection to enhance the understanding of variability in the accident data is discussed in Chapter 8 but is, for the purposes of this research, a source of uncertainty.

2.2.2 Uncertainty in exposure models

Uncertainty has many forms:

- Input uncertainty: missing data and error in measurements lead to uncertainty in the input variables. Input uncertainty can be categorised into four areas:

  Proxy data can be used as an estimate if the actual quantity of exposure needed is undefined or unmeasurable. The relationship between the proxy data and actual exposure is uncertain.

  Random error is a general independent deviation from the population

mean due to measurement error or test conditions.

Systematic error may be due to a calibration error or inaccuracies in assumptions resulting in a systematic bias to the overall result.

Dependence and correlation may exist when there is more than one uncertain event and there is a dependence between the errors in the events.

- Model uncertainty: a model is often a simplification of a complex process. Model uncertainty introduces several issues:

  Structure of the model depends on technical assumptions such as the appropriate statistical distribution required to represent the data. Different assumptions can be modelled and compared to assess the robustness of the results to differing assumptions.

  Validation of the model is important for determining how well the model represents the data. In interpreting the results we must consider whether there are sufficient data to assume a causal relationship, or whether quantities are more likely to be correlated.

  Extrapolation of the model results to other regions of a parameter space may not be appropriate.

  Resolution of data should be considered when the purpose and desired accuracy of the model is decided. Any aggregation of data introduces uncertainty. This type of uncertainty can often be assessed by running the model on a more (or less) disaggregated set of data.

Often uncertainty in a model presents a combination of the issues listed above, and all of these should be considered when attempting to account for uncertainty.

As the exposure data used in this thesis are almost all based on sample surveys, input uncertainty is highly likely. There are published estimates of the uncertainty in the National Traffic estimates (Department for Transport 2010), however other sources of data (for example, the induced exposure data) used in this research do not have these associated uncertainty estimates and are deemed to be substantially more uncertain; therefore these published estimates of uncertainty do not provide much information. Proxy uncertainty is of more concern – an ideal measure of exposure for single vehicle accidents is the amount of time spent on the road network, however we have a measure of traffic in vehicle kilometres. It is suggested that these are compatible, however measures of the number of registered vehicles and other exposure measures used for disaggregation are not. For multi-vehicle accidents the concept is slightly different – as there must be more than one vehicle present for there to be a multi-vehicle accident, the associated ideal measure of exposure must be time when there is more than one vehicle present at any point on the road. The link between this and traffic is less obvious and requires more research in the future.

Kelly and Smith (2009) simply state that if you do not know the exposure exactly then you should assign it a probability distribution. A Bayesian analysis is inferred here and this is a popular approach in Probabilistic Risk Assessment (PRA), due to the flexibility it provides in evaluating multiple exposure scenarios, common in epidemiology studies. For example, Martz and Picard (1995) use a Bayesian approach for expressing uncertainty in Poisson event counts and exposure time in PRA in nuclear power stations and Sohn et al. (2004) evaluate the uncertain exposure to trichloroethylene using a Bayesian model. Bayesian modelling, as discussed in Chapter 3, requires the specification of a prior distribution – this often contains knowledge from previous experiments or theoretical reasoning about the results. Martz and Picard (1995) show that the strength of the prior is particularly important when data are uncertain – a weak prior increases the effect of ignoring

the uncertainties. In PRA, most commonly a Gamma or log-Normal distribution is used for positive parameters and a Normal distribution for unrestricted parameters (Kelly and Smith 2009) as these are good candidates for expressing uncertainty. Specifying the prior distribution introduces model structure uncertainty.

In fact Kelly and Smith (2009), in a recent review, recommend a hierarchical Bayes approach with multistage priors, which has, they state, only become practical with the introduction of the MCMC software BUGs (Lunn et al. 2000). In practice, most studies use non-hierarchical MC simulation which is simple and convenient to use (Hart et al. 2003).

In Chapter 4 we test a series of models using a Bayesian approach and conclude that the most appropriate model is based on a log-Normal distribution with Normal and Gamma priors.

Sohn et al. (2004), in their epidemiological study evaluating the measurement of exposure to trichloroethylene demonstrate the difficulty in estimating exposure from uncertain data. This research used a Bayesian model in a controlled experiment where a group of men were exposed, in the same room, to a fixed level of poison in air. Subsequent measurements of their blood showed two distinct groups with different levels of poison in their blood stream. Sohn et al. (2004) conclude that any risk analysis requires sufficient and reliable information, and suggest that as uncertainty can, in theory, be reduced it is always worth considering if there are any additional exposure data likely or practical to be obtained.

We use three different exposure datasets, in order to disaggregate the exposure into smaller groups. These additional datasets cannot reduce the amount of uncertainty in the exposure measure as they measure different concepts. More certain traffic data on minor roads (traffic estimates for major roads are already based on many sampling points) could, in theory, be collected however this is a huge task and, given the scope of this research, we propose would be of very little benefit.

2.3 Combining information

Often a single exposure data set is of insufficient quality or detail to draw robust conclusions and it is necessary to consult and combine further sources of data to strengthen the ability to draw conclusions. Data can be combined from similar and compatible sources such as information from all automatic traffic counters distributed across the British major road network which provides an overall measure of traffic across the major road network. Alternatively, common in health studies are combinations of broad measures of exposure combined with a small survey of detailed data collection.

Hart et al. (2003) and Molitor et al. (2006) combine a small detailed data set with a reduced set of data over a larger sample. Hart et al. (2003) discuss uncertainty in human food contamination and food consumption when carrying out risk analysis of contaminants and additives in food. They recognise that previous research using detailed exposure data from diet diaries from a small number of individuals results in skewed results, particularly when evaluating the effect of contaminants on disaggregated groups, such as certain age groups or across different regions. Hart et al. (2003) derive human dietary exposure data from a combination of a small sample of detailed diet diaries and a set of simplified data from a medium sized sample. They discuss the use of worst case scenario estimates to take account of the remaining uncertainty in the exposure data, but have achieved a reduced uncertainty due to the combination of two sources of data.

Molitor et al. (2006) estimate the residential exposure to traffic pollution to evaluate the effect on lung function using, similarly to Hart et al. (2003), two sources of data with different levels of detail: continuous long term central site measurements in multiple communities from the Children's Health study and two seasonal short term household level measurements. Previous research has shown that within

community variability may affect health more than between community variability and therefore the more in-depth measurements made within households are important for reducing uncertainty. With the help of a multilevel model using Bayesian MCMC and these two sources of data the authors have developed an improved exposure model which estimates missing measurements and projects data from the short term household level survey to long term exposure whilst encompassing variation in results across communities.

The combination of unrelated sources, such as the three sources we use in this research to improve the disaggregation in the data is less common – in fact we have found little evidence of this being used as a technique to improve exposure data in the scientific literature and the combination of unrelated exposure datasets for detailed accident rate modelling has not been done before.

2.4 Conclusions

Uncertainty (an incomplete set of knowledge) and variability (natural variation in a quantity) are essential concepts to include in a model if we are not to suggest overconfidence in our results. In our exposure data we have both of these and we deal with each differently.

Variability in traffic exposure data comes from road users' choice of vehicle, time of day of driving, road type and demographics. We deal with these sources of variability by combining a series of datasets together enabling us to disaggregate into smaller groups of road users with less variable exposures. Any remaining variability within these groups is treated as a source of uncertainty. There is no previous research which combines data of these types.

Each of the combined datasets present sources of uncertainty. The overall measure of traffic that is commonly used in accident rate calculations is a proxy for acci-

dent exposure and the inputs into this proxy are based on estimates derived from sample surveys across the road network introducing a further level of uncertainty. Combining the datasets together using a model introduces further uncertainty as we assume the data fit certain statistical distributions. We use a Bayesian modelling procedure to represent all of these sources of uncertainty explicitly. There is little previous research which even considers uncertainty in traffic exposure data, and those which exist restrict their analysis to subsets of data where exposure is known and certain.

Chapter 3

Bayesian inference

3.1 Introduction

Bayesian inference presents conclusions about unknown parameters or unobserved data using probability statements, which are conditional on specified known data.

To make these probability statements, a joint probability distribution $p(\theta, y)$ is required for the observed data y, and θ, the unknown parameters. Using conditional probability rules and Bayes' theorem this joint probability distribution can be used to present the required posterior distribution $p(\theta \mid y)$:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

where $p(y)$ is the normalising constant $\sum_\theta p(\theta)p(y \mid \theta)$ or $\int_\theta p(\theta)p(y \mid \theta)d\theta$ for continuous θ.

This fundamental equation, which can be represented as

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

contains the basic components for all Bayesian inference: $p(\theta \mid y)$ is the posterior distribution which gives the probability distribution of the unknown parameters given the data that have been observed; $p(y \mid \theta)$ is the likelihood: the probability of some observed data given some parameter values; and $p(\theta)$ is defined as the prior distribution, the subjective uncertainty about the unknown parameters before any data are observed. The prior distribution of the parameters could be based on previous experiments, theoretical reasoning or subjective assessment. As the quantity of evidence increases, the posterior distribution becomes less dependent on the subjective nature of the prior beliefs.

The basic structure of Bayesian inference described above is very flexible. In particular, simulation techniques discussed in Section 3.2 allow inference on complex multivariate distributions similar to those that we have here.

3.2 Markov chain Monte Carlo

Many multivariate probability distributions are impossible or at least inefficient to sample from. Markov chain Monte Carlo (MCMC) is a method of simulating from such a distribution using Monte Carlo integration[1] to construct a Markov chain which (given satisfied criteria) results in dependent samples from the distribution. This method is particularly important in Bayesian inference because the object of inference is typically a multivariate probability distribution specified as a collection of conditional distributions.

In general an MCMC algorithm generates a sequence of dependent observations θ_i from a normalised density $f(\theta) = g(\theta)/\int g(\theta)d\theta$, starting from an arbitrary θ_0, such that θ_{i+1} is independent of $\theta_{i-1}, \theta_{i-2} \ldots$ given the preceding value θ_i. If the

[1]Monte Carlo integration evaluates $E(f(X))$ by drawing samples $\{X_t : t = 1, \ldots n\}$ from $f(x)$ and then approximating, by the ergodic theorem (essentially a MC version of the law of large numbers) $E(f(X)) = \frac{1}{n}\sum f(X_t)$

appropriate conditions are satisfied, the generated distribution of θ_i converges to the stationary distribution and all θ_i can be assumed to come from the stationary distribution once i is sufficiently high and the chain has converged.

3.2.1 History

Robert and Casella (2010) have produced a comprehensive history of MCMC research which is summarised in this section.

Much of the MCMC research was developed in the 1990s and acceptance of the technique developed dramatically during this time, however some earlier work also deserves recognition. Metropolis et al. (1953) describe their first ideas which led to the start of the MCMC revolution. They proposed a random walk modification to the Monte Carlo method (repeated random sampling) for evaluating complex multi-dimensional integrals with applications in molecular physics. This involved proposing a new value from any (symmetric) probability distribution, called the proposal distribution, and deciding whether to accept the value as a member of the distribution of interest, based on an acceptance probability. Hastings (1970) developed this first MCMC algorithm (later named the Metropolis algorithm) so that the proposal distribution did not have to be symmetric and defined general acceptance probabilities along with warnings about low acceptance rates and difficulties in assessing errors.

In the early 1970s Hammersley and Clifford (1971) developed the concept of specifying joint distributions as a combination of conditional distributions. Geman and Geman (1984) named this technique Gibbs sampling from the study of Gibbs random fields and introduced it into the area of statistical application. However, it was not until Gelfand and Smith (1990) wrote their paper stating that simulating from the joint distribution is the same as simulating from the conditionals (in the limit) that the statistical community became interested in using

MCMC methods. A raft of research on applications manifested quickly as the new opportunities with these methods became clear.

Robert and Casella (2010) suggest that Tierney (1994) produced the most influential paper on the theory of MCMC, including the definitions of assumptions and properties such as convergence of ergodic averages[1] and central limit theorems[2]. Convergence (see Section 3.6) quickly became a property of interest, and in particular defining convergence and speed of convergence. Mengersen and Tweedie (1996) started the research into how fast the algorithms converge.

The reversible jump MCMC algorithm (Green 1995) developed the area by allowing chains to choose models and parameters simultaneously, defining a model choice application of MCMC.

3.3 Gibbs sampler

The Gibbs sampler is one of the simplest MCMC samplers and can be applied when a joint probability distribution which is difficult to sample from can be written as a product of known conditional distributions over each variable. The joint distribution can be simulated via the conditional distributions. For example, if $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is a set of random variables with conditional distributions $p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ then the joint distribution of θ given data y can be derived from

$$p(\theta_1, \theta_2, \ldots, \theta_n \mid y) = \prod_i p(\theta_i \mid y, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$$

[2]A function $h : X \to R$ of a Markov chain $\{X_n\}$ which is irreducible and aperiodic and with stationary distribution $f(.)$ satisfies the Central Limit Theorem ($\sqrt{(}n)$-CLT) if there exists $\sigma^2 < \infty$ such that $n^{-1/2} \sum_{i=1}^{n} [h(X_i) - f(h)] \overset{weakly}{\longrightarrow} N(0, \sigma^2)$

The iterative algorithm, first introduced by Tanner and Wong (1987) in the application of data augmentation, contains an initialisation step followed by a repeated iteration step until convergence appears to have been reached.

1. Initialise: Choose a set of initial values $\theta^{(0)}$
2. Iteration j: Sample each variable from its conditional distribution given the current values of all other variables
   (a) Sample $\theta_1^{(j)}$ from $p(\theta_1^{(j)} \mid y, \theta_2^{(j-1)}, \ldots, \theta_n^{(j-1)})$
   (b) ...
   (c) Sample $\theta_i^{(j)}$ from $p(\theta_i^{(j)} \mid y, \theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_n^{(j-1)})$
   (d) ...
   (e) Sample $\theta_n^{(j)}$ from $p(\theta_n^{(j)} \mid y, \theta_1^{(j)}, \ldots, \theta_{n-1}^{(j)})$
3. Continue step 2 until convergence and beyond for stationary distribution.

3.4 Metropolis-Hastings sampler

If conditional distributions are not available or not easily sampled from, then an arbitrary proposal distribution can be used to sample from and convergence to the required distribution is reached using an alternative algorithm: the Metropolis Hastings algorithm.

Metropolis et al. (1953) proposed the following method:

1. For a set of random variables, set initial values for random variable θ: $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)})$ and set iteration counter to $j = 1$.
2. Move θ from previous position $\theta^{(j-1)}$ according to symmetric proposal distribution such as $q(\theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta-\phi)^2}$ centred at these points to obtain new proposal ϕ for $\theta^{(j)}$.
3. Calculate the acceptance probability of the move α, as shown in equation 3.1 below.

4. If ϕ is not accepted then define $\theta^{(j)} = \theta^{(j-1)}$, else define $\theta^{(j)} = \phi$.
5. Repeat from step 2 until convergence and beyond for stationary distribution.

This was generalised by Hastings (1970) to remove the necessity for $q(\cdot, \cdot)$, the proposal distribution, to be symmetrical. The generalised version requires the proposal distribution $q(\theta, \phi)$ to satisfy the detailed balance equations $p(\theta)q(\theta, \phi) = p(\phi)q(\phi, \theta)$ for all (θ, ϕ), where $p(\cdot, \cdot)$ is the stationary distribution. The acceptance probability is defined as

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{p(\phi)/q(\theta, \phi)}{p(\theta)/q(\phi, \theta)} \right\} \tag{3.1}$$

There are many different choices of $q(\theta, \phi)$ which lead to different algorithms. Three of the most common choices of algorithms are described below.

Metropolis algorithm (Metropolis et al. 1953)

For symmetric proposals, $q(\phi, \theta) = q(\theta, \phi)$ and the acceptance probability simplifies to

$$\alpha(\theta, \phi) = \min \left\{ \frac{p(\phi)}{p(\theta)}, 1 \right\}$$

Random walk algorithm

The random walk algorithm generates a new proposal state based on a perturbation of the previous state: $\phi = \theta + \epsilon$ where ϵ is independent of θ. If ϵ is symmetric around 0 then, as above, $q(\phi, \theta) = q(\theta, \phi)$ and

$$\alpha(\theta, \phi) = \min \left\{ \frac{p(\phi)}{p(\theta)}, 1 \right\}$$

The random walk algorithm tends to take a long time to explore the whole space.

Independence sampler

The independence sampler generates a new state which is independent of the previous state: $q(\phi \mid \theta) = q(\phi)$. The acceptance probability is defined as

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{p(\phi)/q(\phi)}{p(\theta)/q(\theta)} \right\}$$

This sampler tends to work either quite well or particularly badly. It is recommended that the proposal and posterior distributions are similar in order for this sampler to converge more quickly, and the tails of the proposal distribution should dominate those of the posterior density in order to get sufficient extreme values from the chain (Bolstad 2011).

3.4.1 Tuning the transition probabilities

In general, the choice of the proposal distribution $q(\theta, \phi)$ from any of the samplers above comes from a family of distributions which have scale and/or spread parameters that need to be defined. The choice of these parameters affects the number of times a move is accepted and the coverage of the chain over the sample space; for example, when using a Normal distribution as the proposal for a random walk algorithm, a smaller variance will result in higher acceptance rates but less coverage of the parameter space. For a random walk process with Normal distributions, acceptance probabilities of around 45% for one-dimensional problems and 23% for high dimensional problems should be used to tune the spread/scale parameters. (Roberts et al. 1997)

3.5 Optimization

The improvement of these MCMC samplers is sometimes necessary to ensure efficient sampling and convergence within a reasonable run time.

3.5.1 Implementation

There are three options when sampling N items from the posterior distributions:

- sample N chains and take the r^{th} value from each chain. These are independent and rN iterations are required;
- use the ergodic theorem to sample N from 1 chain after burn-in b. These are not independent but only $b + N$ iterations are required;
- sample every k^{th} item in chain after burn-in b. These are quasi-independent if k is big enough[3], and $b + kN$ iterations are required.

The general consensus is that using N different chains is computationally inefficient and unnecessary. If convergence is quick then only one chain should be needed. Sufficient spacing between states should be chosen for a pseudo independent sample (Smith and Roberts 1991).

3.5.2 Optimization of Gibbs algorithm

Scanning strategies

Different strategies for updating the values in each iteration are optimal in different circumstances. The most commonly used updating scans are the deterministic scan and the random scan described below.

- update each component in order at each iteration: *deterministic scan*

[3]One way of determining a reasonable k is to observe the autocorrelation values, choosing a k value which suitably reduces the autocorrelation (Tierney 1994)

- visit each component in order and then in reverse order at the following iteration
- update a randomly drawn component in each iteration
- update all components in a random order: *random scan*
- each component visited every k^{th} iteration

Reparameterisation

An iteration moves along the co-ordinate axis of the θ_i. If the components of θ are weakly dependent then the space is covered quickly. If some of the components of θ are highly correlated then the chain moves are small and convergence is slow. Reparameterisation may lead to better convergence times. There are no rules to determine suitable transformations, however, a linear transformation which results in a diagonal covariance matrix usually works well.

Blocking

Instead of component-wise moving in each iteration, components can be combined into blocks and general moves can be made across the space, taking into account dependences between components. Blocks can be updated in a random order or in a set order (this leads to a sampler which does not satisfy the detailed balance equation). This generally leads to a more mobile sampler (Andrieu et al. 2003). The joint full conditional for each block of parameters must be known to determine the moves.

The general rule is to block as much as possible such that the joint full conditionals are easy to sample from. It is possible to update components in more than one block.

3.6 Convergence

Convergence of the chain is based on two values: burn in 'b' – the number of iterations that are required before the chain becomes independent of the starting value θ_0; and length of chain after burn in 'm' - the total number of iterations required to assume the components represent the stationary distribution $p(.)$ after burn in.

A number of schemes for evaluating convergence fall into two groups:

- theoretically (formal): this is difficult to obtain and apply to practical problems;
- statistically (informal): this can show non-convergence, but never guarantees convergence.

No known scheme guarantees convergence, so implementing as many as possible is advised.

3.6.1 Informal convergence monitors

A series of plots and summary statistics can be used to show non convergence informally:

1. Plot times series path (or summary stats for high dimensional space) – when the central path is not remaining in the same area this suggests non-convergence;
2. Plot ergodic averages of 1st t values in Markov chain and compare them to later averages – major differences in these averages will suggest non-convergence;
3. Compare histograms of a set of t points with a further set of t points further

along the chain (beware of metastable chains[4]) – different histograms will highlight non-convergence;

4. For Metropolis Hastings compute the average percentage of iterations for which moves are accepted – if the percentage is low, movement within the space may be restricted and convergence may not have yet been reached.

3.6.2 Formal Convergence methods

Time Series

It is possible to assess whether convergence can be safely assumed to hold by comparing the ergodic averages ($\bar{\theta}$) of two sections of the chain, a and b (Geweke 1992):

$$\frac{\bar{\theta_a} - \bar{\theta_b}}{\sqrt{\hat{var}(\theta_a) + \hat{var}(\theta_b)}} \rightarrow N(0,1)$$

where $\hat{var}(\theta_a)$ is the variance of the values in section a of the chain.

Multiple chains

Alternatively, Gelman and Rubin (1992) suggest starting several (C) chains from different points, and testing whether the dispersion σ_θ^2 within the chains is different to the dispersion between the chains. This can be achieved formally by defining B as the variance between the chains and W as the within variance. Define θ_j^c as the j^{th} iteration of the c^{th} chain.

$$B = \frac{m}{C-1} \sum_{c=1}^{C} (\bar{\theta}_c - \bar{\theta})^2$$

[4]Metastable chains are stationary for a finite period of time only

$$W = \frac{1}{C(m-1)} \sum_{c=1}^{C} \sum_{j=1+b}^{m} (\theta_j^c - \bar{\theta}^c)^2$$

where m is the length of each chain after burn in (of length b).

Under convergence, σ_θ^2 can be consistently estimated from $\hat{\sigma}_\theta^2$: the weighted average of B and W:

$$\sigma_\theta^2 = \frac{m-1}{m} W + \frac{1}{m} B$$

Define the potential scale reduction measure R as

$$\hat{R} = \sqrt{\frac{\hat{\sigma}_\theta^2}{W}}$$

which tends to 1 as $m \to \infty$ and Gelman and Rubin (1992) suggest accepting convergence if $R < 1.2$.

Starting several chains with different initial values will also check for meta-stability in the chains.

Conditional distributions

Finally, assume θ can be divided into two blocks: θ_1 and θ_2. The difference criterion tends to 0 for all θ if the chain has converged.

Difference criterion: $\hat{\eta} = p(\theta_1 \mid \theta_2)\hat{p}(\theta_2) - p(\theta_2 \mid \theta_1)\hat{p}(\theta_1)$

where $\hat{p}$ is the marginal distribution. Alternatively, the ratio criteria ξ_1 and ξ_2 will be close on convergence.

Ratio convergence:

$$\hat{\xi_1} = \frac{p(\theta_2 \mid \theta_1)\hat{p}(\theta_1)}{p(\theta_2^\star \mid \theta_1^\star)\hat{p}(\theta_1^\star)}$$

and

$$\hat{\xi}_2 = \frac{p(\theta_1 \mid \theta_2)\hat{p}(\theta_2)}{p(\theta_1^\star \mid \theta_2^\star)\hat{p}(\theta_2^\star)}$$

where $\theta^\star = (\theta_1^\star, \theta_2^\star)'$ is another value from the state space.

3.7 Missing data

Missing or censored data are relatively easy to deal with within the theory of MCMC algorithms. The idea is that the missing data are treated as unknowns along with the model parameters. Gibbs sampling can be used to solve this augmented data problem.

Given $z = (y, y')$ where y is the known data and y' is the unknown data, the unknowns for the Gibbs sampler are now θ and y' leading to the conditionals: $p(\theta \mid y', y) = p(\theta \mid z)$ and $p(y' \mid \theta, y) = p(y' \mid \theta)$ where the former is the conditional which occurs when there is no missing data and the latter is the sampling distribution, under the model, of y' given θ.

For censored data, the results are similar: treat the censored observations as unknowns and the corresponding full conditionals are the joint posterior for θ which is the same as that which would have been observed with no censoring, and the second is the joint distribution of the censored observations given θ (Smith and Roberts 1991).

3.8 Software

The Bayesian inference Using Gibbs Sampling (BUGs) software is designed to apply MCMC techniques to complex statistical models. WinBUGs is a version of the BUGs software written by researchers at MRC Biostatistics Unit at Cambridge University (Lunn et al. 2000) which has a user interface which allows users to draw

or 'doodle' their model in pictures. WinBUGs has been used in Chapters 5 and 6.

3.9 Bayesian Model selection

3.9.1 Introduction

Model selection is the process of finding the best fitting, least complex model given the data. There are many different ways of selecting models, from a range of different information criteria (usually comprising the log maximum likelihood and a penalty for complexity) to a Bayesian probabilistic approach, which compares models using the posterior model probabilities (from likelihoods marginalised over the parameters). The Bayesian approach for a finite set of models is discussed in more detail below.

We define the set of models as $M = \{M_1, \ldots, M_K\}$, the associated set of parameters for each model as θ_k and the observed data as y.

The probability that a chosen model is the 'correct' model is given by the posterior model probability:

$$p(M_k \mid y) = \frac{p(y \mid M_k)p(M_k)}{\sum_k p(y \mid M_k)p(M_k)}$$

where

$$p(y \mid M_k) = \int p(y \mid \theta_k, M_k)p(\theta_k \mid M_k)\,\mathrm{d}\theta_k$$

$p(M_k)$ is the model prior probability, $p(y \mid \theta_k, M_k)$ is the marginal likelihood and $p(\theta_k \mid M_k)$ is the prior probability of the parameters given the model. The model and parameter prior probabilities need to be specified, using subjective or uninformative priors. Some comments on choosing model priors have been included below. The marginal likelihood values are generally difficult to calculate

directly, but a number of approximate methods exist, such as Gelfand and Dey's estimator (Gelfand and Dey 1994), bridge sampling (Meng and Wong 1996) and Laplace approximation (Tierney and Kadane 1986, described in Section 3.9.3) where the likelihood can be assumed to be approximately Normal.

The model which results in the highest posterior model probability is generally selected. Pairwise comparisons of models can be carried out using Bayes Factors, described below.

3.9.2 Model priors

A popular prior for models is the uniform prior $p(M_k) = 1/K$: the posterior model probability calculation above reduces to a constant times the marginal likelihood and this prior favours all models equally. It is, as such, noninformative. However, as Chipman et al. (2001) shows, it is often not noninformative on model characteristics such as model size – giving higher weight to models of the most common number of parameters. A number of alternatives for noninformative priors have been suggested, such as the improper prior proposed by Jeffreys (1961) for nested models which applies equal probabilities for those models with equal dimensions $p(M_k^{(d)}) = \frac{1}{d+1}$ where $d = 0, 1, \ldots$ is the dimension of the model.

3.9.3 Laplace approximation

Laplace approximation (Tierney and Kadane 1986) allows asymptotic approximation of marginal posterior densities and is used in model selection for approximating the marginal likelihood. The context is the evaluation of $I(\beta) = \int e^{h(\beta)}\,\mathrm{d}\beta$ and the overall process can be summarised as expanding the integral using Taylor series, disregarding negligible terms and normalising (Gill 2002).

In Section 3.9.1, we are required to evaluate $\int p(y \mid \theta_k, M_k)p(\theta_k \mid M_k)\,\mathrm{d}\theta_k$. If we

let $p(y \mid \theta_k, M_k)p(\theta_k \mid M_k) = g(\beta) = \exp[h(\beta)]$ and $\hat{\beta} = \text{argmax } g(\beta)$ (we assume that the posterior density is highly peaked around $\hat{\beta}$, the posterior mode), then using a Taylor series expansion,

$$h(\beta) = h(\hat{\beta}) + h'(\hat{\beta})^T(\beta - \hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^T h''(\hat{\beta})(\beta - \hat{\beta}) + \dots$$

which approximates to a normal pdf which can be evaluated using standard forms.

If $\hat{\beta} = \text{argmax } \beta$ then $h'(\hat{\beta}) = 0$ by definition which results in

$$\begin{aligned}
\int g(\beta)d\beta &\simeq \int \exp\left[h(\hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^T h''(\hat{\beta})(\beta - \hat{\beta})\right] \mathrm{d}\beta \\
&= g(\hat{\beta}) \int \exp\left[\frac{1}{2}(\beta - \hat{\beta})^T h''(\hat{\beta})(\beta - \hat{\beta})\right] \mathrm{d}\beta \\
&= g(\hat{\beta})|-h''(\hat{\beta})|^{-1/2}(2\pi)^{d/2}
\end{aligned}$$

where $d = \dim(\beta)$.

If the sample size is sufficiently large[5] then a first order expansion provides an adequate result. In fact the relative error of this first order approximation is $O(n^{-1})$ which, when the approximation is applied to the numerator and the denominator, reduces to $O(n^{-2})$ (Kass and Raftery 1995).

3.9.4 Bayes Factors

Bayes factors can be used to compare two models directly. Gill (2002) shows that the posterior odds ratio in favour of Model 1 over Model 2 is produced from Bayes law:

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1)/p(y)}{p(M_2)/p(y)} \times \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)}$$

$$\text{Posterior odds} = \text{Prior odds} \times \text{Bayes factor}$$

[5]Kass and Raftery (1995) suggest that their 'rough feeling' is a sample size of 5 times the dimension of β is 'worrisome' and 20 times is sufficiently large

from which the Bayes factor is therefore defined as

$$B_{12}(y) = \frac{p(M_1 \mid y)/p(M_1)}{p(M_2 \mid y)/p(M_2)}$$

which reduces to the common likelihood ratio if $p(M_1) = p(M_2) = \frac{1}{2}$. These factors can be computed using the methods detailed in Section 3.9.1 such as Laplace approximation described in Section 3.9.3.

There are no explicit rejection or acceptance thresholds implicit in Bayes factors however several authors have provided advice on the interpretation of such factors, including Kass and Raftery (1995) who provide categories for the 'weight of evidence' (Good 1985) against model 2 compared to model 1 as shown in Table 3.1.

Table 3.1: Kass and Raftery (1995) weight of evidence for Bayes Factor interpretation

$2\log(B_{12})$	B_{12}	Evidence against M_2
0 to 2	1 to 3	Not worth more than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
> 10	> 150	Decisive

Bayes factors are more flexible than their classical counterparts. It is possible to compare multiple hypotheses without correction factors and it is not necessary for the models being compared to be nested.

3.9.5 Model averaging

A process of model selection chooses the single most likely model which represents the data. This strategy ignores any uncertainty associated with the choice of a model. Model averaging avoids this problem by taking a number of reasonable models with their associated posterior probabilities and producing an average

(weighted) result.

In general, the posterior density of a parameter θ when model averaging over a series of models $\{M_j \text{ for } j \text{ in } 1:K\}$ is given by

$$p(\theta \mid y) = \sum_{j=1}^{K} p(M_j \mid y) p(\theta \mid M_j, y)$$

where $p(\theta \mid y)$ is the weighted average of conditional posterior densities, $p(\theta \mid M_j, y)$.

In the context of MCMC model selection algorithms, model averaging is simple: parameter inference is defined as an average of the parameter posterior distributions for all or a selection of high probability models.

3.10 Model checking

At the end of any statistical analysis it is important to check the model. Gelman et al. (1996) and Green et al. (2009) state that classical goodness of fit tests are not appropriate for many models, such as complex probabilistic or discrete response models, especially if there are restrictions on the parameters, probabilistic constraints or it is difficult to put the model into a standard statistical form, as these tests rely on knowing a reference sampling distribution. Bayesian model checking is more flexible and should be based on three aspects (Gelman et al. 1996):

1. sensitivity of inference to reasonable change in prior distribution and likelihood;
2. plausible posterior inferences given substantive context of model; and
3. fit of the model for the data.

Points 1 and 2 are fairly self explanatory and require a sensitivity analysis on the chosen prior and likelihood or subjective review of inference given the context, respectively. This section concentrates on step 3.

The substantive question is also different in a Bayesian context: we do not assess whether the model is correct (as it almost always is not) but ask if it is reasonable that the data have arisen by chance given the model (Gelman et al. 1996). If not, then Gelman et al. (2004) pose a further question which allows us to consider whether the deficiencies are sufficient to adapt the model: do the model deficiencies and simplifying assumptions have a noticeable effect on the substantive inferences?

Green et al. (2009) discuss three options for assessing model fit when standard methods such as residual plots are not appropriate: cross validation, external validation and posterior predictions. Cross validation partitions the data into a number of subsets, defines the model on one of the subsets and uses the remaining subsets to compare with predictions from the model. This method has been referred to as the gold standard but is computationally intensive. External validation uses all the data to produce a model and makes predictions about future data from the model. Future data must then be collected to assess the fit and therefore this method is only viable in certain situations. Posterior prediction uses posterior distributions generated from the model to compare to the observed data in order to assess discrepancies in the model. This method appears to be the preferred method: Lynch and Western (2004) show that it is flexible to a range of models, explicitly accounts for parameter uncertainty as the distributions are taken directly from the posterior distributions, and it is possible to derive p-values for evaluating the probability that the data arose by chance; however, Green et al. (2009) suggest that it is likely to be over optimistic in the assessment of model fit as each data point is used to generate the model.

The discrepancies used to assess model fit can be flexible and should measure relevant features of the model. In a posterior predictive check, Gelman et al. (2004) show the chosen discrepancies can be measured using a test statistic T from which it is possible to directly summarise discrepancies between data generated from the model y^{rep} and observed data y. If $T(y)$ is not contained well within the empirical distribution of $T(y^{rep})$ then the model is found to be lacking. It is also useful to examine graphical representations of the discrepancy measures or results from the posterior predictive simulations to assess where the model is failing to represent the observed data.

If the model is a good fit then data generated from the model should look similar to the observed data and the discrepancies should be small. The location of $T(y)$ within the distribution of $T(y^{rep})$ directly relates to the evaluation of a p-value: the probability that replicated data could be more extreme than the observed data (Gelman et al. 2004). If $T(y)$ is located towards the extremes of the distribution of $T(y^{rep})$ then a small p-value indicates that the observed data is unlikely to be replicated if the model is true.

3.11 Summary

Many of the techniques discussed in this chapter are employed in Chapters 4, 5 and 6 to evaluate accident exposure and risk in a probabilistic manner.

Chapter 4

Exposure modelling

Traffic flow, defined as the number of vehicles kilometres travelled in Great Britain, is used to monitor trends in road travel across the country on different road types, at different times of day and year, and by different vehicle types. These trends help to define areas of congestion and inform government expenditure for road construction, improvement and structural maintenance.

In addition, traffic flows are also used as exposure data to estimate the risk of being involved in an accident. Typically accident risk is defined as the number of accidents by the number of vehicle kilometres travelled each year which can be disaggregated as much as the traffic flow data allows: that is, by region, road type, time of day and year, and vehicle type. Flow information about different types of car is not available and other data must be substituted. In general, traffic flow data are replaced by information on the number of registered vehicles each year in order to compute accident rates in these situations. af Wahlburg and Dorn (2007) suggest that traffic flows cannot be directly derived from the distribution of the different car types registered, as drivers of different car types generally have different driving habits. Therefore a method of estimating the traffic flow

distribution of different car types is required.

A Bayesian approach to modelling flow data has been used. This approach allows the transparent incorporation of prior information and provides not only flow estimates but associated measures of uncertainty. In this chapter a number of different model formulations are described, helping to develop the final model, along with some results of each model.

4.1 Data

The sources of data used in this chapter are described in Sections 1.3.1, 1.3.2 and 1.3.4. The variable names used throughout are given below:

- x_{+yr}: number of car kilometres travelled each year (y) on each road type (r) (Table 1.1)
- z_{cy}: number of registered cars by year (y) and car type (c) (Table 1.2)
- e_{cr}: proportion of cars involved, but not at fault, in accidents by car type c within road types r (Table 1.3)
- x_{cyr}: number of vehicle kilometres travelled each year (y) on each road type (r) by each car type (c) (to be estimated)

Additional variables and parameters are described where appropriate.

The data that are available give us some idea of the relationship between year and road type, year and car type, and distribution of car types on each road. We do not have data that can inform us of the distribution of road use by individual car types, or of any changing distribution of different road use by different car types by year. We have therefore had to make a simplifying assumption. We have assumed that the distribution of car types using each road type has not changed over the 12 year period.

To illustrate the models, three scenarios have been used:

- a simulated test dataset with three car types, two years and two road types ($C = 3$, $Y = 2$ and $R = 2$) where x_{+yr}, z_{cy} and e_{cr} are specified separately from the true data, and the exposure variable $\boldsymbol{x}$ is derived and known, as shown in Tables 4.7, in order to check the validity of the model;

- a subset of the true data x_{+yr}, z_{cy} and e_{cr} with dimensions $C = 3$, $Y = 2$ and $R = 2$, shown as the margins of Table 4.1. The values of x_{+yr} and e_{cr} have been factored accordingly for this smaller number of car types and x_{cyr} is unknown; and

- the complete set of real data shown in Tables 1.1, 1.2 and 1.3 with dimensions $Y = 12$, $C = 6$ and $R = 3$.

For the second and third scenarios we compute a basic unweighted combination of x_{+yr}, z_{cy} and e_{cr} via an approximate proportional fitting algorithm and the results of the models have been compared with these estimates.

Table 4.1: Known inputs for small exposure datasets e_{cr}, x_{+yr} and z_{cy}

A roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005				64.6
2006				65.4
e_{cr}	0.24	0.20	0.55	
Minor roads	**4x4 & people carriers**	**Large saloons**	**Medium saloons**	x_{+yr}
2005				52.2
2006				52.8
e_{cr}	0.28	0.10	0.63	
z_{cy}	4x4 & people carriers	Large saloons	Medium saloons	
2005	2.4	1.7	5.7	
2006	2.6	1.7	5.5	

4.2 A heuristic proportional fitting algorithm

In order to test that the MCMC computer coding was producing sensible results an alternative method of combining the three data sources was derived for comparison. The method of proportional fitting was applied where the three data sources form the three marginal totals. In a standard iterative proportional fitting algorithm (Deming and Stephan 1940) each marginal total is a sum over a different margin of the same data source. In this thesis the margins are based on different data sources: the three marginal totals (shown for the small dataset in bold in Table 4.2) are $\boldsymbol{x}^{+} = \{x_{+yr} : y = 1, \ldots, Y;\ r = 1, \ldots, R\}$, $\boldsymbol{e} = \{e_{cr} : c = 1, \ldots, C,\ r = 1, \ldots, R\}$ and $\boldsymbol{z} = \{z_{cy} : c = 1, \ldots, C;\ y = 1, \ldots, Y\}$, where $\boldsymbol{e}$ and $\boldsymbol{z}$ are adjusted so that their totals are equivalent to $\boldsymbol{x}^{+}$ (shown as italics in Table 4.2). The starting point is that all values of x_{cyr} are equal and sum to the total of x^{+}.

Table 4.2: Known and derived inputs for proportional fit algorithm (iteration 0) on small exposure dataset

A roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	19.6	19.6	19.6	**64.6**
2006	19.6	19.6	19.6	**65.4**
$\boldsymbol{e_{cr}}$	**0.24**	**0.20**	**0.55**	
e_{cr}	*31.6*	*26.6*	*71.8*	**130.0**

Minor roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	19.6	19.6	19.6	**52.2**
2006	19.6	19.6	19.6	**52.8**
$\boldsymbol{e_{cr}}$	**0.28**	**0.10**	**0.63**	
e_{cr}	*29.1*	*10.1*	*65.7*	**104.9**

$\boldsymbol{z_{cy}}$	4x4 & people carriers	Large saloons	Medium saloons	
2005	**2.4**	**1.7**	**5.7**	
2006	**2.6**	**1.7**	**5.5**	
2005	*28.8*	*20.4*	*68.3*	
2006	*31.2*	*20.4*	*65.9*	**234.9**

In each iteration, the individual x_{cyr} values are proportionally adjusted to equal the distribution of margin $\boldsymbol{e}$ (Table 4.3) followed by margin $\boldsymbol{x}^+$ (Table 4.4), and finally margin $\boldsymbol{z}$ (Table 4.5). This process is repeated until it stabilises.

Table 4.3: Adjustment for e_{cr} in heuristic proportional fit algorithm for exposure data (iteration 1)

A roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	15.8	13.3	35.9	65.0
2006	15.8	13.3	35.9	65.0
Σ_y	31.6	26.6	71.8	130.0
	31.6	*26.6*	*71.8*	

Minor roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	14.5	5.1	32.9	52.5
2006	14.5	5.1	32.9	52.5
Σ_y	29.1	10.1	65.7	104.9
	29.1	*10.1*	*65.7*	

Σ_r	4x4 & people carriers	Large saloons	Medium saloons	
2005	30.3	18.4	68.8	
2006	30.3	18.4	68.8	234.9
2005	*28.8*	*20.4*	*68.3*	
2006	*31.2*	*20.4*	*65.9*	**234.9**

As the three margins come from different datasets the $\boldsymbol{x} = \{x_{cyr} : c = 1, \ldots, C; y = 1, \ldots, Y;\ r = 1, \ldots, R\}$ do not converge to satisfy each of the margins. It is either possible to satisfy $\boldsymbol{z}$ which controls year and car type applying the road distribution from $\boldsymbol{x}^+$ or to satisfy margins $\boldsymbol{x}^+$ and $\boldsymbol{e}$ which control year and road type, and car type within road type respectively. Therefore we end up with two sets of proportionally fitted flow estimates but these are very similar. We choose the estimate satisfying $\boldsymbol{x}^+$ and $\boldsymbol{e}$ to use and this is shown in Table 4.6.

Table 4.4: Adjustment for x^+ in heuristic proportional fit algorithm for exposure data (iteration 1)

A roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	15.7	13.2	35.7	64.6
2006	15.9	13.4	36.1	65.4
Σ_y	31.6	26.6	71.8	130.0
	31.6	*26.6*	*71.8*	

Minor roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	14.5	5.0	32.7	52.2
2006	14.6	5.1	33.1	52.8
Σ_y	29.1	10.1	65.8	104.9
	29.1	*10.1*	*65.7*	

Σ_r	4x4 & people carriers	Large saloons	Medium saloons	
2005	30.1	18.3	68.4	
2006	30.5	18.5	69.2	234.9
2005	*28.8*	*20.4*	*68.3*	
2006	*31.2*	*20.4*	*65.9*	**234.9**

Table 4.5: Adjustment for z_{cy} in heuristic proportional fit algorithm for exposure data (iteration 1)

A roads	4x4 & people carriers	Large saloons	Medium saloons	x_{+yr}
2005	15.1	14.2	36.2	65.5
2006	15.9	14.2	34.9	65.0
Σ_y	31.0	28.5	71.0	130.5
	31.6	*26.6*	*71.8*	
Minor roads	**4x4 & people carriers**	**Large saloons**	**Medium saloons**	x_{+yr}
2005	13.9	5.4	33.1	52.4
2006	14.7	5.4	31.9	52.0
Σ_y	28.6	10.8	65.0	104.4
	29.1	*10.1*	*65.7*	
Σ_r	4x4 & people carriers	Large saloons	Medium saloons	
2005	28.8	20.4	68.3	
2006	31.2	20.4	65.9	234.9
2005	*28.8*	*20.4*	*68.3*	
2006	*31.2*	*20.4*	*65.9*	**234.9**

Table 4.6: Final iteration in heuristic proportional fit algorithm for exposure data

A roads	4x4 & people carriers	Large saloons	Medium saloons
2005	15.1	13.2	36.3
2006	16.5	13.4	35.5
Minor roads	**4x4 & people carriers**	**Large saloons**	**Medium saloons**
2005	13.9	5.0	33.3
2006	15.2	5.1	32.5

4.3 Initial model without induced exposure

Initially, a model was developed that took information from $\boldsymbol{x}^+$ and $\boldsymbol{z}$.

Model

The initial model was defined as a natural first step for introducing some variability into the data. It comprises two regression parts – firstly the registered vehicle data is modelled as a Normal distribution with mean defined by the unknown exposure data $\boldsymbol{x}$, summed over r multiplied by some scalar constant β, and variance defined by a diagonal matrix with diagonal elements defined by τ^2. The second part defines the unknown exposure variable $\boldsymbol{x}$ as Normally distributed with mean defined by $\boldsymbol{\mu}$ and variance defined by a diagonal matrix with diagonal elements defined by σ^2. The parameters β, $\boldsymbol{\mu}$, τ and σ can be defined from the start or given prior distributions.

$$\boldsymbol{z} \mid \boldsymbol{x} \sim N(\beta A\boldsymbol{x}, \tau^2 \boldsymbol{I}_{CY})$$
$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_{CYR})$$

where $\boldsymbol{A}$ is a $CY \times CYR$ sum matrix over r, $\boldsymbol{z}$ is a $C \times Y$ vector, $\boldsymbol{x}$ is a $C \times Y \times R$ vector, $\boldsymbol{\mu}$ is a $C \times Y \times R$ vector and β, τ and λ are scalars.

The method of completing the square was used to derive the posterior parameters for the required $\boldsymbol{x}$.

$$
\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{z}) &\propto p(\boldsymbol{z} \mid \boldsymbol{x}) p(\boldsymbol{x}) \\
&\propto \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{z} - \beta \boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{z} - \beta \boldsymbol{A}\boldsymbol{x}) - \frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu})^T(\boldsymbol{x} - \boldsymbol{\mu}) \right) \\
&\propto \exp\Bigg(-\frac{1}{2}\Big\{ \boldsymbol{x}^T \left[\frac{\beta^2}{\tau^2}\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{I}\sigma^{-2} \right] \boldsymbol{x} \\
&\qquad + \boldsymbol{x}^T \left[-\frac{\beta \boldsymbol{A}^T \boldsymbol{z}}{\tau^2} - \frac{\boldsymbol{\mu}}{\sigma^2} \right] + \left[-\frac{\beta \boldsymbol{z}^T \boldsymbol{A}}{\tau^2} - \frac{\boldsymbol{\mu}^T}{\sigma^2} \right] \boldsymbol{x} \Big\} \Bigg)
\end{aligned}
$$

Therefore

$$
\boldsymbol{x} \mid \boldsymbol{z} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})
$$

where, using $\exp\big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\theta}) \big) \equiv \exp\big(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} \big)$,

$$
\begin{aligned}
\boldsymbol{\Sigma}^{-1} &= \frac{\beta^2}{\tau^2}\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{I}\sigma^{-2} \\
\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} &= \frac{\beta \boldsymbol{A}^T \boldsymbol{z}}{\tau^2} + \frac{\boldsymbol{\mu}}{\sigma^2} \\
\boldsymbol{\theta} &= \left(\frac{\beta^2}{\tau^2}\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{I}\sigma^{-2} \right)^{-1} \left(\frac{\beta \boldsymbol{A}^T \boldsymbol{z}}{\tau^2} + \frac{\boldsymbol{\mu}}{\sigma^2} \right)
\end{aligned}
$$

Only the known variable $\boldsymbol{z}$ is used in the posterior. The other known observed variable $\boldsymbol{x}^+$ is used to constrain the resulting exposure variable $\boldsymbol{x}$.

The constraint works as follows: $\boldsymbol{x}$ is reparameterised as $(\boldsymbol{x}^\star, \boldsymbol{x}^+)^T$ where $\boldsymbol{x}^\star = \{x_{cyr} : c = 1, \ldots C-1; y = 1, \ldots, Y;\ r = 1, \ldots, R\}$ which, together with $\boldsymbol{x}^+$, defines $\boldsymbol{x}$ by $(\boldsymbol{x}^\star, \boldsymbol{x}^+)^T = \boldsymbol{B}\boldsymbol{x}$. $\boldsymbol{B}$ is a square matrix which sums over c for the final $Y \times R$ rows of $\boldsymbol{x}$.

Therefore $(\boldsymbol{x}^\star, \boldsymbol{x}^+ \mid \boldsymbol{z}) \sim N(\boldsymbol{B}\boldsymbol{\theta}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T)$ can be manipulated using the standard Multivariate Normal theory: in general, if $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as

$$
\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}
$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}$$

then, the distribution of $\boldsymbol{X_1}$ conditional on $\boldsymbol{X_2} = \boldsymbol{a}$ is Multivariate Normal $(\boldsymbol{X_1} \mid \boldsymbol{X_2} = \boldsymbol{a}) \sim N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ where

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

which leads us to

$$\begin{aligned}(\boldsymbol{x}^\star \mid \boldsymbol{x}^+, \boldsymbol{z}) \sim N\Big(&\boldsymbol{\mu}_{x^\star} + \boldsymbol{\Sigma}_{x^\star x^+}\boldsymbol{\Sigma}_{x^+x^+}^{-1}(\boldsymbol{x}^+ - \boldsymbol{\mu}_{x^+}), \\ &\boldsymbol{\Sigma}_{x^\star x^\star} - \boldsymbol{\Sigma}_{x^\star x^+}\boldsymbol{\Sigma}_{x^+x^+}^{-1}\boldsymbol{\Sigma}_{x^+x^\star}\Big)\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{x^\star} &= \boldsymbol{B\theta}_{[1:((C-1)\times Y\times R),\ ]} \\
\boldsymbol{\mu}_{x^+} &= \boldsymbol{B\theta}_{[((C-1)\times Y\times R+1):(C\times Y\times R),]} \\
\boldsymbol{\Sigma}_{x^\star x^\star} &= \boldsymbol{B\Sigma B}^T{}_{[1:((C-1)\times Y\times R),1:((C-1)\times Y\times R)]} \\
\boldsymbol{\Sigma}_{x^\star x^+} &= \boldsymbol{B\Sigma B}^T{}_{[1:((C-1)\times Y\times R),((C-1)\times Y\times R+1):(C\times Y\times R)]} \\
\boldsymbol{\Sigma}_{x^+ x^\star} &= \boldsymbol{B\Sigma B}^T{}_{[((C-1)\times Y\times R+1):(C\times Y\times R),1:((C-1)\times Y\times R)]} \\
\boldsymbol{\Sigma}_{x^+ x^+} &= \boldsymbol{B\Sigma B}^T{}_{[((C-1)\times Y\times R+1):(C\times Y\times R),((C-1)\times Y\times R+1):(C\times Y\times R)]}
\end{aligned}$$

A direct simulation from equation 4.1 is possible (if β, $\boldsymbol{\mu}$, τ and σ are pre-specified) using the Cholesky decomposition of $\boldsymbol{\Sigma}$. Figure 4.1 shows the results of this model on the small datasets with inputs $\beta = 11$, $\tau = 20$ and $\sigma = 2$ with $\boldsymbol{\mu}$ equal to the proportional fit described in Section 4.2, compared to the proportional fit of the exposure data (Traffic flow (PF)).

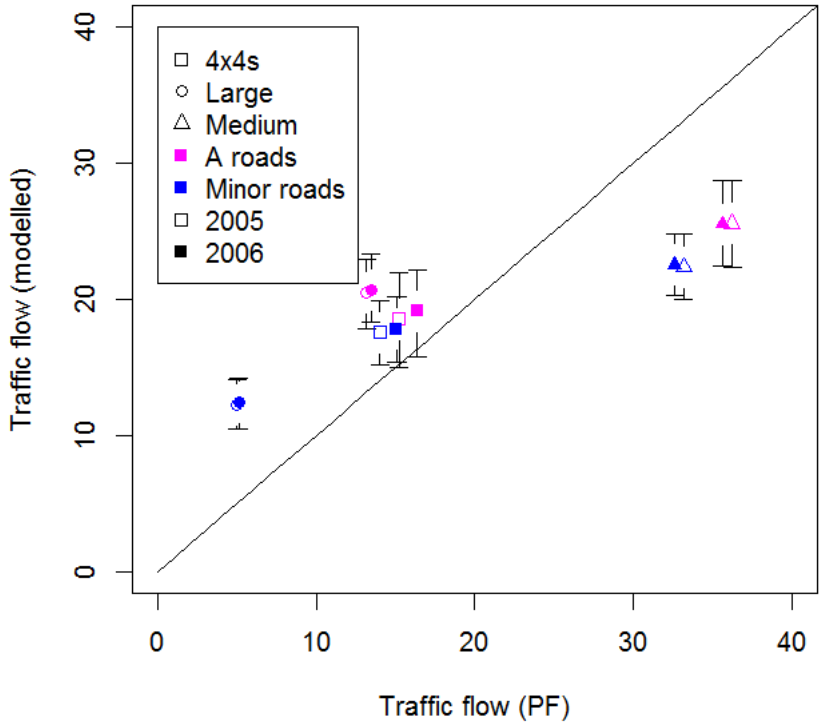


Figure 4.1: Modelled exposure (traffic flow – modelled) x_{cyr} against proportional fit exposure (traffic flow – PF) x_{cyr} for small data in model without induced exposure data

The results of this model rely heavily on the deterministic inputs. In addition, as σ^{-2} tends to infinity, $\boldsymbol{\Sigma}$ becomes singular and $\boldsymbol{\Sigma}_{x^+x^+}$ becomes non-invertible. An alternative model with non-deterministic inputs for β and $\boldsymbol{\mu}$ generated from the posteriors shown in Appendix B.1.1 removes the need to invert $\boldsymbol{\Sigma}_{x^+x^+}$.

Define

$$\boldsymbol{x} = \boldsymbol{D}\begin{pmatrix}\boldsymbol{x}^\star \\ \boldsymbol{x}^+\end{pmatrix}$$

where $\boldsymbol{D}$ is $\boldsymbol{B}^{-1}$, and the posterior distribution becomes

$$p(\boldsymbol{x}^\star \mid \boldsymbol{z}, \boldsymbol{x}^+) \propto \exp\left(-\begin{pmatrix}\boldsymbol{x}^\star \\ \boldsymbol{x}^+\end{pmatrix}^T \left[\frac{\beta^2}{2\tau^2}\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{D} + \frac{1}{2\sigma^2}\boldsymbol{D}^T\boldsymbol{D}\right]\begin{pmatrix}\boldsymbol{x}^\star \\ \boldsymbol{x}^+\end{pmatrix} + \begin{pmatrix}\boldsymbol{x}^\star \\ \boldsymbol{x}^+\end{pmatrix}^T \left[\frac{\beta}{\tau^2}\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{z} + \frac{1}{\sigma^2}\boldsymbol{D}^T\boldsymbol{\mu}\right]\right)$$

A Metropolis-Hastings sampler described in Section 3.4 would be required for this alternative.

Conclusions

This initial model formulation was defined to add variability into the data structure. Computationally it is uncomplicated – not requiring MCMC techniques until σ was large; however, it is not a natural model for causal inference. Once MCMC techniques are required the benefits of this simple model are removed. A more complicated model is required as this basic model assumes that the distribution of traffic across different road types is the same for all car types. That is, there is no information in the model which describes the interaction between c and r.

4.4 Introducing induced exposure

We found no evidence to suggest that equal distributions of different car types across road types was a fair assumption, and in fact the opposite is proposed: af Wahlburg and Dorn (2007) suggest that drivers of different car types generally have different driving habits, including use of different road types. This includes larger cars such as 4x4s and people carriers thought to be being used more regularly for long journeys on primarily main roads, and small cars generally being used for short journeys on local routes.

It is difficult to source data which identifies the distribution of different car types on different road types, so induced exposure techniques, described in Section 1.3.3 have been used. This additional data source $\boldsymbol{e}$ estimates the distribution of different car types c on different road types r. These data are normalised over r due to the data collection procedure.

In this section we consider two regression models which include induced exposure data.

4.4.1 Truncated Normal model

The second model considers the exposure measure $\boldsymbol{x} = \{x_{cyr} : c = 1, \ldots, C; y = 1, \ldots, Y; r = 1, \ldots, R\}$ to have a truncated Multivariate Normal distribution (as a MVN model allows negative values for $\boldsymbol{x}$) with the mean defined by the product of the number of registered vehicles by year and car type $(\boldsymbol{z})$[1] and an unknown parameter $\boldsymbol{\beta} = \{\beta_{cr} : c = 1, \ldots, C; r = 1, \ldots, R\}$. Its variance τ^2 is an additional unknown parameter. The sum of traffic over car type c $(\boldsymbol{x}^+)$ is used in limiting the posterior for $\boldsymbol{x}$ but is not defined in the model.

$$\begin{aligned} \boldsymbol{x} &\sim tN(\boldsymbol{\beta z}, \tau^2) \qquad x_{cyr} > 0 \\ \boldsymbol{e} &\sim N(\boldsymbol{\alpha z}^+ \boldsymbol{\beta}, \lambda^2) \end{aligned} \tag{4.2}$$

The second part of this model (4.2) models the normalised induced exposure data $\boldsymbol{e}$. This is modelled as a Multivariate Normal distribution[2] with a mean defined by the product of $\boldsymbol{z}^+ = \{z_{c+} : c = 1, \ldots, C\}$ where $z_{c+} = \sum_y z_{cy}$ and two unknown

[1] Throughout this chapter where vectors of unmatching lengths are shown multiplied together these vectors have been augmented with repeated values. Here, $\boldsymbol{\beta z} = (\boldsymbol{I}_R \otimes \boldsymbol{z}_{diag})(\boldsymbol{\beta} \otimes \boldsymbol{1}_Y)$ and $\boldsymbol{\alpha z}^+ \boldsymbol{\beta} = (\boldsymbol{\alpha} \otimes \boldsymbol{z}^+)^T \boldsymbol{\beta}$ where $\boldsymbol{I}_R$ is an identity matrix of dimension $R \times R$, $\boldsymbol{z}_{diag}$ is a diagonal matrix with values of $\boldsymbol{z}$ on the diagonal, $\boldsymbol{1}_Y$ is a vector of 1s of length Y and $\otimes$ is the Kronecker product.

[2] In theory, this part of the model could also be modelled using a truncated Multivariate Normal distribution. In practice this was an unnecessary complication as the derived relative variability was sufficiently small that the posterior distribution did not approach its limits.

parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = \{\alpha_r : r = 1, \ldots, R\}$, and variance λ^2. This error term λ^2 represents the inherent uncertainty around the method of induced exposure, and the limited amount of data that are available. The first parameter $\boldsymbol{\alpha}$ represents information about the different amounts of use of different road types r by cars, and the second parameter $\boldsymbol{\beta}$ is an estimate of the distribution of different car types on different road types, taking into account relevant information from $\boldsymbol{e}$, $\boldsymbol{z}$ and $\boldsymbol{x}^+$.

This model is not mathematically tractable because we only observe $\boldsymbol{x}^+$, not $\boldsymbol{x}$, so the observed data likelihood is not of standard form. Therefore we apply MCMC simulation techniques to estimate posterior distributions from the priors and likelihoods. The chosen prior distributions for the unknown parameters are assumed independent and defined as the normal conjugate distributions: $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ priors are Normally distributed, and the variance parameters have Inverse Gamma (IG) prior distributions. The priors are designed to be fairly uninformative.

$$
\begin{aligned}
&\beta_{cr} \sim N(\beta_0, \sigma_\beta^2) \qquad c = 1, \ldots, C,\ r = 1, \ldots, R \\
&\alpha_r \sim N(\alpha_0, \sigma_\alpha^2) \qquad r = 1, \ldots, R \\
&\lambda^2 \sim IG(\lambda_a, \lambda_b) \\
&\tau^2 \sim IG(\tau_a, \tau_b)
\end{aligned}
$$

The joint posterior distribution, given the model and the prior distributions, is given in equation 4.3. It is possible to derive the individual posterior distributions for the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, τ^2 and λ^2, and the unknown flow $\boldsymbol{x}$ from which to sample, and these are shown in equations (4.4) - (4.7).

$$
p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau^2, \lambda^2 \mid \boldsymbol{z}, \boldsymbol{e}) = p(\boldsymbol{x} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{z}) p(\boldsymbol{e} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda^2) p(\boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\tau^2) p(\lambda^2) \quad (4.3)
$$

As $\boldsymbol{x}$ is distributed with a truncated Normal distribution, there exists an extra normalising constant in the joint posterior which carries through to the posteriors for $\boldsymbol{\beta}$ and τ, making them non conjugate.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{x}, \tau^2, \lambda^2, \boldsymbol{\alpha}, \boldsymbol{e}, \boldsymbol{z}) \propto \exp \Bigg\{ & -\frac{1}{2\tau^2} \sum_{cyr} (x_{cyr} - \beta_{cr} z_{cy})^2 \\ & -\frac{1}{2\lambda^2} \sum_{cr} (e_{cr} - \alpha_r z_{c+} \beta_{cr})^2 \\ & -\frac{1}{2\sigma_\beta^2} \sum_{cr} (\beta_{cr} - \beta_0)^2 \Bigg\} \\ & \prod_{cyr} \frac{1}{1 - \Phi(-\frac{\beta_{cr} z_{cy}}{\tau})} \end{aligned} \tag{4.4}$$

where $\Phi(.)$ is the Normal distribution function.

$$\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \lambda^2, \boldsymbol{e}, \boldsymbol{z} \overset{ind}{\sim} N(\boldsymbol{\phi}, \boldsymbol{\Omega}) \tag{4.5}$$

where $\boldsymbol{\phi}$ is a $r \times 1$ dimensional vector and $\boldsymbol{\Omega}$ is a $r \times r$ matrix such that:

$$\begin{aligned} \phi_r &= \left(\frac{\sum_c (z_{c+} \beta_{cr})^2}{\lambda^2} + \frac{1}{\sigma_\alpha^2} \right)^{-1} \left(\frac{\sum_c e_{cr} z_{c+} \beta_{cr}}{\lambda^2} + \frac{\alpha_0}{\sigma_\alpha^2} \right) \\ \Omega_r &= \left(\frac{\sum_c (z_{c+} \beta_{cr})^2}{\lambda^2} + \frac{1}{\sigma_\alpha^2} \right)^{-1} \end{aligned}$$

$$\begin{aligned} p(\tau^2 \mid \boldsymbol{x}, \boldsymbol{\beta}, \tau_a, \tau_b, \boldsymbol{z}) \propto & \left(\frac{1}{2\pi\tau^2} \right)^{\frac{CYR}{2}} \left(\frac{1}{\tau^2} \right)^{\tau_a + 1} \exp \Big\{ -\frac{1}{2\tau^2} \sum_{cyr} (x_{cyr} - \beta_{cr} z_{cy})^2 - \frac{\tau_b}{\tau^2} \Big\} \\ & \prod_{cyr} \frac{1}{1 - \Phi(-\frac{\beta_{cr} z_{cy}}{\tau})} \end{aligned} \tag{4.6}$$

$$\lambda^2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_a, \lambda_b, \boldsymbol{z} \sim IG \left(\frac{CR}{2} + \lambda_a, \frac{1}{2} \sum_{cr} (e_{cr} - \alpha_r z_{c+} \beta_{cr})^2 + \lambda_b \right) \tag{4.7}$$

Similarly to the constraint in the initial model without induced exposure, the posterior distribution for the three dimensional exposure parameter $\boldsymbol{x}$ is based on the first part of the model together with the additional aggregated known traffic flow information from $\boldsymbol{x}^+$.

The full parameter $\boldsymbol{x}$ can be defined as a combination of a subset of itself up to row $(C-1) \times Y \times R$ (called $\boldsymbol{x}^\star$) combined with the known aggregated traffic flow data $\boldsymbol{x}^+$ via a matrix $\boldsymbol{B}$:

$$\boldsymbol{B}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}^\star \\ \boldsymbol{x}^+ \end{pmatrix}$$

where $\boldsymbol{B}$ is a square $C \times Y \times R$ matrix such that the first $C-1 \times Y \times R$ rows are the identity matrix followed by $Y \times R$ rows of 1s and 0s summing over each car type. The posterior can then be defined as

$$\begin{pmatrix} \boldsymbol{x}^\star \\ \boldsymbol{x}^+ \end{pmatrix} \sim N(\boldsymbol{B}\boldsymbol{\gamma}, \tau^2 \boldsymbol{B}\boldsymbol{B}^T)$$

where $\boldsymbol{\gamma}$ is the CYR vector $\boldsymbol{\gamma} = \boldsymbol{\beta}\boldsymbol{z}$. Using standard Multivariate Normal theory once again, the required conditional distribution can be defined as

$$\begin{aligned} \boldsymbol{x}^\star \mid (\boldsymbol{x}^+, \boldsymbol{z}) \sim N(&\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}^+ - \boldsymbol{\mu}_2), \\ &\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}_1 &= \boldsymbol{B}\boldsymbol{\beta}\boldsymbol{z}[1 : (C-1)YR] \\ \boldsymbol{\mu}_2 &= \boldsymbol{B}\boldsymbol{\beta}\boldsymbol{z}[(C-1)YR + 1 : CYR] \\ \boldsymbol{\Sigma} &= \tau^2 \boldsymbol{B}\boldsymbol{B}^T = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{aligned}$$

A Gibbs sampling strategy with Metropolis-Hastings steps for $\boldsymbol{\beta}$ and τ has been implemented below.

Simulated study

To illustrate the models we generate a test dataset with three car types, two years and two road types ($C = 3$, $Y = 2$ and $R = 2$). With this test data we know the inputs and outputs and can test the model by comparing the outcomes from the model with the known values.

In order to generate the test data. $\boldsymbol{\beta}$, τ and λ have been defined. $\boldsymbol{\beta}$ has been set as

$$\boldsymbol{\beta} = \begin{pmatrix} 200 \\ 300 \\ 100 \\ 150 \\ 400 \\ 600 \end{pmatrix} \tag{4.8}$$

which has been specified independently from the defined $\boldsymbol{z}$ data. τ and λ have been specified as 5 and 0.5 respectively, for the purposes of adding random noise to these components.

Values of the induced exposure measure $\boldsymbol{e}$, parameter $\boldsymbol{\alpha}$, unknown flow $\boldsymbol{x}$ and its aggregated vector $\boldsymbol{x}^+$ were computed from the model form shown in 4.2 and perturbed with this random noise. The exposure values x_{cyr} are shown in Table 4.7.

We take only the information from the test data that we would know in a real data situation, that is $\boldsymbol{e}$, $\boldsymbol{z}$ and $\boldsymbol{x}^+$. We set uninformative priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0 = 1$, σ_β^2 and $\sigma_\alpha^2 = 1\,000$. A number of different priors for τ^2 and λ^2 have been specified which result in similar outcomes. Two choices are shown

Table 4.7: Derived exposure values for test study on truncated Normal exposure model

Year	Road type	Car1	Car2	Car3
1	1	529	184	2143
	2	778	278	3686
2	1	485	158	2713
	2	786	259	3932

in Table 4.8 and displayed in the results (Tight and Diffuse).

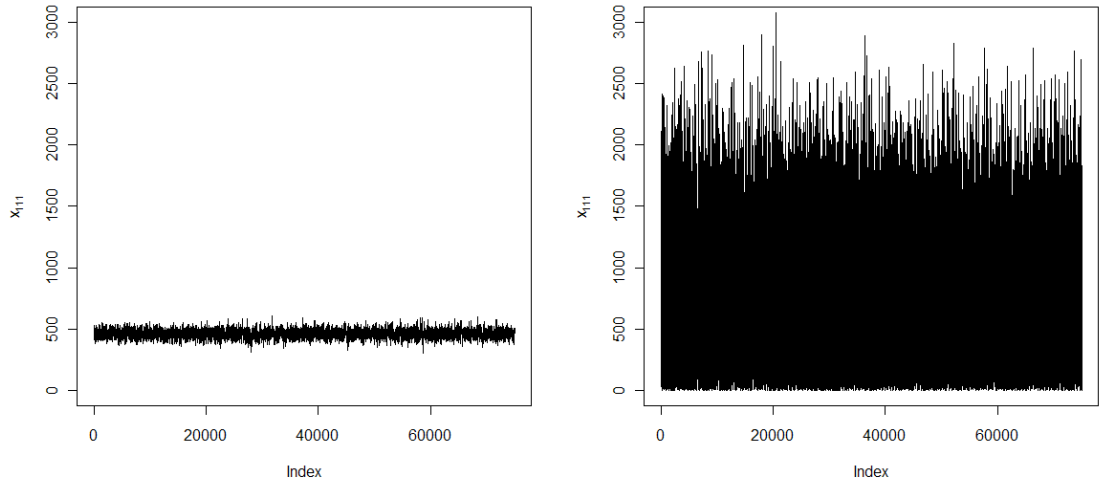
Table 4.8: Priors for τ^2 and λ^2 in simulated study on truncated Normal exposure model

Parameter	Tight	Diffuse
τ_a	4	4
τ_b	40	300
λ_a	4	4
λ_b	0.004	0.03

These relate to expected variances for $\boldsymbol{x}$ and $\boldsymbol{e}$ of 13 and 0.001 for Tight (a precise prior) and 100 and 0.01 in Diffuse (an imprecise prior).

The simulation is run 100 000 times and the first 25 000 runs are removed for burn in. Time series plots, shown in Figure 4.2 for the first component of $\boldsymbol{x}$, appear to show good convergence over the number of iterations. A series of different starting points led to similar results.

The mean $\boldsymbol{x}$ predictions have been compared with the test $\boldsymbol{x}$ values shown in Table 4.7 above in Figure 4.3, and the other known parameters are compared to the model predictions in Table 4.9. Modelled and test results should be similar and points on the graph should sit approximately in a straight line. Results of the model in Figure 4.3 are similar – with the less precise priors producing wider confidence intervals, and more precise values giving better estimates for the parameter values in Table 4.9.

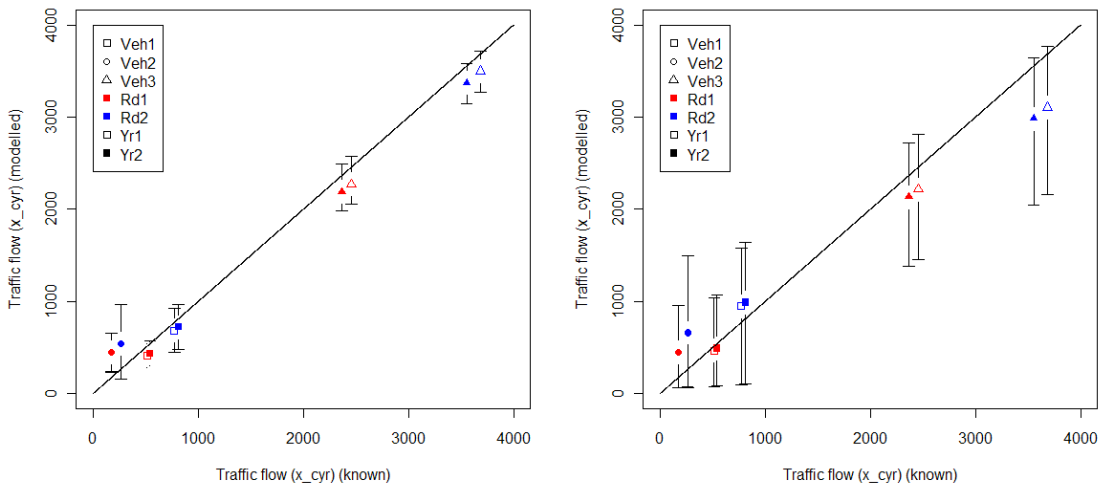


(a) Tight (b) Diffuse

Figure 4.2: Time series of unknown exposure for 4x4s in 1999 on Motorways (x_{111}: iterations 25 000 to 100 000) for simulated data in truncated Normal model for tight and diffuse priors

Table 4.9: Modelled and actual β_{cr} and α_r posterior mean parameter values for simulated data on truncated Normal exposure model

Parameter	Car	Road	Tight	Diffuse	Known test
β	1	1	167	167	200
β	1	2	307	318	300
β	2	1	227	270	100
β	2	2	205	241	150
β	3	1	376	363	400
β	3	2	581	565	600
α	-	1	1.6×10^{-4}	1.6×10^{-4}	1.6×10^{-4}
α	-	2	1.0×10^{-4}	1.0×10^{-4}	1.1×10^{-4}
τ	-	-	7	9	50
λ	-	-	0.06	0.09	0.50



(a) Tight (b) Diffuse

Figure 4.3: Modelled exposure (traffic flow – modelled) x_{cyr} against actual exposure (traffic flow – known) x_{cyr} for simulated data in truncated Normal exposure model with tight and diffuse priors

4.4.2 Truncated Normal model without road type parameter α

The previous model (model 4.2) contains the parameter $\boldsymbol{\alpha} = \{\alpha_r : r = 1, \ldots, R\}$ which is defined as representing information about the different amounts of use of road type r by cars, and assumed to be independent of $\boldsymbol{e}$, the induced exposure data. The information that we have about road and car type interaction has to be normalised by road type and therefore $\boldsymbol{\alpha}$ cannot be informed by data. This model (model 4.9) removes the need for the parameter $\boldsymbol{\alpha}$ and models $\boldsymbol{g}$, the multivariate logit of $\boldsymbol{e}$. $\boldsymbol{g}$ is a vector of length $R \times (C-1)$ and is the log of the ratio of e_{ri} from $i = 1, \ldots C-1$ and e_{rC} where $r = 1, \ldots R$, assumed to be Normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\epsilon^2 \boldsymbol{K}$. $\boldsymbol{K}$ is a diagonal matrix with off-diagonal component defined here as 0.5 to infer some dependence between elements of μ.

$$\begin{aligned} \boldsymbol{x} &\sim N(\boldsymbol{\beta z}, \tau^2) \\ \boldsymbol{g} &\sim N(\boldsymbol{\mu}, \epsilon^2 \boldsymbol{K}) \end{aligned} \tag{4.9}$$

where

$$\begin{aligned} g_{ri} &= \log \frac{e_{ri}}{e_{rC}} \\ \mu_{ri} &= \log \frac{\beta_{ri} z_i}{\beta_{rC} z_C} \\ K_{ij} &= \begin{cases} 1.0 & \text{if } i = j \\ 0.5 & \text{if } i \neq j \end{cases} \end{aligned}$$

and $i, j = 1, \ldots C - 1$

The joint posterior (equation 4.10) has a similar form to the initial truncated Normal model (model 4.3) with $\boldsymbol{e}$ replaced with the logit model $\boldsymbol{g}$, $\boldsymbol{\alpha}$ removed and the variance term λ^2 exchanged for ϵ^2. Changing the second part of the model only affects the posterior distribution of $\boldsymbol{\beta}$ (shown in 4.11), and introduces the posterior distribution for ϵ (equation 4.12) with similar format to that of λ in equation 4.7 where the prior for ϵ^2 is of the form $\epsilon^2 \sim IG(\epsilon_a, \epsilon_b)$.

$$p(\boldsymbol{x}, \boldsymbol{g}, \boldsymbol{\beta}, \tau^2, \epsilon^2 \mid \boldsymbol{z}, \boldsymbol{e}) = p(\boldsymbol{x} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{z}) p(\boldsymbol{g} \mid \boldsymbol{\beta}, \epsilon^2, \boldsymbol{z}, \boldsymbol{e}) p(\boldsymbol{\beta}) p(\tau^2) p(\epsilon^2) \tag{4.10}$$

$$p(\boldsymbol{\beta} \mid \tau^2, \boldsymbol{x}, \boldsymbol{z}, \epsilon^2, \boldsymbol{g}, \boldsymbol{\mu}) = \{2\pi\tau^2\}^{-\frac{CYR}{2}} \exp\left\{-\frac{1}{2\tau^2}\sum_{cyr}(x_{cyr} - \beta_{cr}z_{cy})^2\right\}$$
$$\{(2\pi\epsilon^2)^R|K|\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\epsilon^2}\sum_{r}\sum_{i,j=1}^{C-1} {K_{ij}}^{-1}(g_{ri} - \mu_i)(g_{rj} - \mu_j)\right\}$$
$$\{2\pi\sigma_\beta^2\}^{-\frac{CR}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2}\sum_{cr}(\beta_{cr} - \beta_0)^2\right\} \tag{4.11}$$

$$\epsilon^2 \mid \beta_{cr}, \epsilon_a, \epsilon_b, \boldsymbol{z}, \boldsymbol{e} \sim IG\left(\frac{R}{2} + \epsilon_a, \frac{1}{2}\sum_{r}\sum_{i,j=1}^{C-1}(g_{ri} - \mu_i)[K^{-1}]_{ij}(g_{rj} - \mu_j) + \epsilon_b\right) \tag{4.12}$$

A Gibbs sampling strategy with Metropolis-Hastings parts is used to sample from the posterior distribution, with blocking over r used to generate β. Blocking, discussed in Section 3.5.2, is used to update homogeneous sections of the data individually so that the sampler moves more easily through the space.

Simulated study

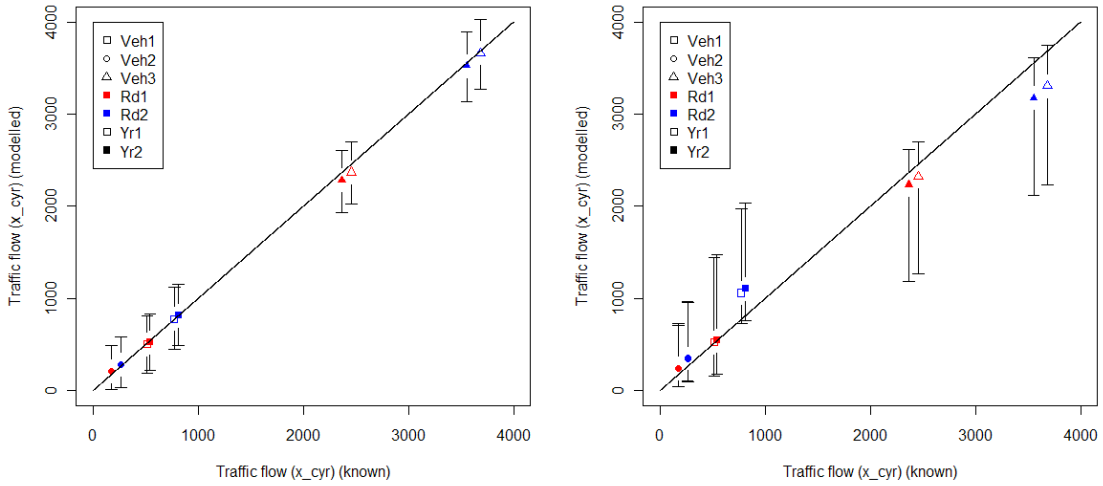
The test data shown in Table 4.7 with three car types, two years and two road types ($C = 3$, $Y = 2$ and $R = 2$) have been used here as in Section 4.4.1 for demonstration purposes. Priors on $\boldsymbol{\beta}$ and τ^2 remain the same. Priors on ϵ^2 have been generated from the prior information given on λ^2 in Table 4.8 resulting in a precise prior with a mean of 0.015 ($\epsilon_a = 3$, $\epsilon_b = 0.03$) and an imprecise prior with a mean of 0.5 ($\epsilon_a = 4$, $\epsilon_b = 1.5$) for ϵ^2.

The simulation is run 100 000 times and the first 25 000 runs are removed for burn in. Good convergence was observed over the number of iterations and a series of different starting points led to similar results. Table 4.10 shows the results of the

model for the $\boldsymbol{\beta}$ parameters compared to the known test data, and Figure 4.4 compares the modelled and known test $\boldsymbol{x}$ values for the tight and diffuse priors.

Table 4.10: Modelled and actual β_{cr} parameter values for simulated data on truncated Normal exposure model with α_r removed

Parameter	Car	Road	Tight	Diffuse	Known test
β	1	1	222	199	200
β	1	2	291	411	300
β	2	1	103	120	100
β	2	2	159	193	150
β	3	1	378	378	400
β	3	2	601	538	600



(a) Tight (b) Diffuse

Figure 4.4: Modelled exposure (traffic flow – modelled) x_{cyr} against actual exposure (traffic flow – known test data) x_{cyr} for test data on truncated Normal model with α_r removed over tight and diffuse priors

4.4.3 Conclusions

The two models, 4.2 and 4.9, use the truncated Normal distribution to model exposure. Two different sets of priors, an informative and a non-informative set,

produce similar results with more accurate parameter posterior means and smaller confidence intervals in the precise prior.

4.5 Log-Normal model

The natural next step from a truncated Normal distribution is a log-Normal model. Some of the benefits over a truncated Normal model are that the posteriors are restricted to positive real numbers in a more natural way than with a truncation method, errors become proportionate rather than additive, and that, in this case, the posterior distributions are conjugate. For these reasons the log-Normal distribution is commonly used for exposure models (Cullen and Frey 1999).

A similar regression model to the truncated Normal model, described in Section 4.4.1, is considered here. Once again, this uses the three datasets discussed in Section 1.3, and aims to estimate the three dimensional array $\boldsymbol{x} = \{x_{cyr} : c = 1, \ldots, C;\ y = 1, \ldots, Y;\ r = 1, \ldots, R\}$ which represents traffic in vehicle kilometres disaggregated by year y, road type r and car type c.

Model form

Our final model considers the exposure measure $\boldsymbol{x}$, over car type c, year y and road type r, to have a Multivariate log-Normal distribution, with the mean defined by the sum of the log of the number of registered vehicles by year and car type ($\boldsymbol{z} = \{z_{cy} : c = 1, \ldots, C;\ y = 1, \ldots, Y\}$) and an unknown parameter $\boldsymbol{\beta} = \{\beta_{cr} : c = 1, \ldots, C;\ r = 1, \ldots, R\}$ which remains constant over time. Its variance τ^2 is an additional unknown parameter. What is observed is $\boldsymbol{x}^+ = \{x_{+yr} : y = 1, \ldots, Y;\ r = 1, \ldots, R\}$, the sum of traffic over car type c and it is this which is used in formulating the posterior for $\boldsymbol{x}$.

The second part models the log of the normalised induced exposure data $\boldsymbol{e} =$

$\{e_{cr} : c = 1, \ldots, C,\ r = 1, \ldots, R\}$ which gives information about the types of cars (c) travelling on different road types (r). Once again, we use a continuous distribution to approximate a discrete sampling process. This second Multivariate log-Normal distribution has a mean defined by the sum of $\log(z_{c+})$ and two unknown parameters, and variance λ^2. This error term λ^2 represents the inherent uncertainty around the method of induced exposure, and the limited amount of data that are available.

$$\begin{aligned} \log \boldsymbol{x} &\sim N(\boldsymbol{\beta} + \log \boldsymbol{z}, \tau^2) \\ \log \boldsymbol{e} &\sim N(\boldsymbol{\alpha} + \log \boldsymbol{z}^+ + \boldsymbol{\beta}, \lambda^2) \end{aligned} \tag{4.13}$$

Similarly to Section 4.4.1, the first parameter $\boldsymbol{\beta}$ models the log of the distribution of different car types on different road types. The second parameter $\boldsymbol{\alpha} = \{\alpha_r : r = 1, \ldots, R\}$ is a nuisance parameter representing information about the different amounts of use of different road types r by cars. The posterior distribution for this parameter is heavily influenced by the relative proportions of different road types in the two counties in which the induced exposure data were collected.

The chosen prior distributions for the unknown parameters are assumed independent and defined as the normal conjugate distributions: $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ priors are Normally distributed, and the variance parameters have Inverse Gamma (IG) prior distributions. The priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are designed to be fairly uninformative. The priors on λ^2 and τ^2 effectively derive weights to the different datasets representing how confident we are about the data. More precise priors lead to

heavier weights for data we are more certain about.

$$\begin{aligned}
&\beta_{cr} \sim N(\beta_0, \sigma_\beta^2) \qquad c = 1, \ldots, C,\ r = 1, \ldots, R \\
&\alpha_r \sim N(\alpha_0, \sigma_\alpha^2) \qquad r = 1, \ldots, R \\
&\lambda^2 \sim IG(\lambda_a, \lambda_b) \\
&\tau^2 \sim IG(\tau_a, \tau_b)
\end{aligned} \tag{4.14}$$

The joint posterior distribution, given the model and the prior distributions, is given in equation 4.15. It is possible to derive the individual conditional posterior distributions for the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, τ^2 and λ^2, shown in equations 4.16 – 4.19, and the unknown exposure $\boldsymbol{x}$.

$$\begin{aligned}
p(\log \boldsymbol{x}, \log \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau^2, \lambda^2 \mid \boldsymbol{z}, \boldsymbol{e}) = p(\log \boldsymbol{x} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{z}) p(\log \boldsymbol{e} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda^2, \boldsymbol{z}) \\
p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\tau^2) p(\lambda^2)
\end{aligned} \tag{4.15}$$

$$\boldsymbol{\beta} \mid \boldsymbol{x}, \tau^2, \lambda^2, \boldsymbol{\alpha}, \boldsymbol{z}, \boldsymbol{e} \sim N(\boldsymbol{\theta}, \boldsymbol{\Delta}) \tag{4.16}$$

where $\boldsymbol{\theta}$ is a vector of length $C \times R$ and $\boldsymbol{\Delta}$ is a matrix of dimensions $CR \times CR$ such that:

$$\begin{aligned}
d_{cr} = \Delta_{cr}^{-1} \theta_{cr} &= \frac{1}{\tau^2} \sum_y \Big(\log x_{cyr} - \log z_{cy} \Big) + \frac{\log e_{cr}}{\lambda^2} - \frac{\log z_{c+}}{\lambda^2} - \frac{\alpha_r}{\lambda^2} + \frac{\beta_0}{\sigma_\beta^2} \\
\Delta_{cr}^{-1} &= \frac{y}{\tau^2} + \frac{1}{\lambda^2} + \frac{1}{\sigma_\beta^2} \\
\theta_{cr} &= \Delta_{cr} \times d_{cr}
\end{aligned}$$

$$\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \lambda^2, \boldsymbol{z}, \boldsymbol{e} \sim N(\boldsymbol{\phi}, \boldsymbol{\Omega}) \tag{4.17}$$

where $\boldsymbol{\phi}$ is a vector of length R and $\boldsymbol{\Omega}$ is a matrix of dimension $R \times R$ such that:

$$\begin{aligned} m_r = \Omega_r^{-1}\phi_r &= \frac{1}{\lambda^2}\sum_c \Big(\log e_{cr} - \log z_{c+} - \beta_{cr} \Big) + \frac{\alpha_0}{\sigma_\alpha^2} \\ \Omega_r^{-1} &= \frac{c}{\lambda^2} + \frac{1}{\sigma_\alpha^2} \\ \phi_r &= \Omega_r \times m_r \end{aligned}$$

The conditional posterior distributions for the two variance parameters τ^2 and λ^2 are Inverse Gamma distributions.

$$\tau^2 \mid \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{z}, \tau_a, \tau_b \sim IG\left(\frac{CYR}{2} + \tau_a, \frac{1}{2}\sum_{cyr} \Big\{ \log x_{cyr} - (\log z_{cy} + \beta_{cy}) \Big\}^2 + \tau_b \right) \tag{4.18}$$

$$\lambda^2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{e}, \boldsymbol{z}, \lambda_a, \lambda_b \sim IG\left(\frac{CR}{2} + \lambda_a, \frac{1}{2}\sum_{cr} \Big\{ \log e_{cr} - (\log z_{c+} + \alpha_r + \beta_{cr}) \Big\}^2 + \lambda_b \right) \tag{4.19}$$

The posterior distribution for the three dimensional exposure measure $\boldsymbol{x}$ is simulated using MCMC. Each parameter is updated in turn ($\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, τ^2 and λ^2) using a Gibbs sampling strategy in a deterministic scan. The vector $\boldsymbol{x}$ is updated last using Metropolis Hastings. The proposal distribution is based on a random walk algorithm (see Section 3.4) where the perturbation is Normally distributed with mean 0 and variance specified such that the acceptance probabilities are around 20%. At each iteration the perturbation is constrained such that the known aggregated traffic flow information from $\boldsymbol{x}^+$ is maintained.

Results

To illustrate this model we use all three datasets: a simple test dataset with three car types, two years and two road types ($C = 3$, $Y = 2$ and $R = 2$); a small set of real data on the same dimensions, shown in Table 4.1; and the complete set of real data with $Y = 12$, $C = 6$ and $R = 3$ discussed in Section 1.3. For each of these real datasets the proportional fitting results described in Section 4.2 have been compared with these estimates.

Simulated study

A small set of test data were defined as discussed in Section 4.4.1. Here we define $\boldsymbol{\beta}$ as the logarithm of the βs defined in equation 4.8.

$$\beta_{cr} = \begin{pmatrix} 5.30 \\ 5.70 \\ 4.61 \\ 5.01 \\ 5.99 \\ 6.40 \end{pmatrix} \tag{4.20}$$

τ^2 and λ^2 have been specified as 0.01, thus weighting both parts equally, for the purposes of adding random noise to these components.

The simulated test data for $\boldsymbol{x}$ are shown in Table 4.11.

We take the simulated values of $\boldsymbol{e}$, $\boldsymbol{x}^+$ and $\boldsymbol{z}$ and apply them to the model with a range of priors. The results shown below are based on the following priors:

Table 4.11: Derived exposure values for test study on truncated Normal exposure model

Year	Road type	Car1	Car2	Car3
1	1	515	176	2450
	2	772	263	3686
2	1	536	168	2375
	2	811	260	3558

$$\begin{aligned} \tau^2 &\sim IG(5, 0.5) \\ \lambda^2 &\sim IG(3, 0.1) \\ \boldsymbol{\beta} &\sim N(0, 1000) \\ \boldsymbol{\alpha} &\sim N(0, 1000) \end{aligned} \tag{4.21}$$

The MCMC is run for 100 000 steps and the first 25 000 iterations are removed for burn in.

In Figure 4.5 we have compared the mean model predictions for $\boldsymbol{x}$ with the known values shown in Table 4.11. Results for the other parameters are shown in Table 4.12. If the computation is effective then we would expect that the modelled and actual results will be very similar and points on the graph would sit along the diagonal as each dataset is equally influential. Error bars in Figure 4.5 are 95% posterior intervals. A series of different starting points led to similar results.

The time series plots in Figure 4.6 of the MCMC samples (after burn-in) for a selection of the modelled parameters appear to show good convergence over the number of iterations.

Table 4.12: Modelled and known test β_{cr}, α_r, τ and λ parameter values for simulated study in log-Normal exposure model

Parameter	Car	Road	Modelled	Known test
β	1	1	5.33	5.30
β	1	2	5.79	5.70
β	2	1	4.74	4.61
β	2	2	5.00	5.01
β	3	1	5.94	5.99
β	3	2	6.35	6.40
α	-	1	-8.72	-8.68
α	-	2	-9.13	-9.12
τ	-	-	0.31	0.10
λ	-	-	0.21	0.10

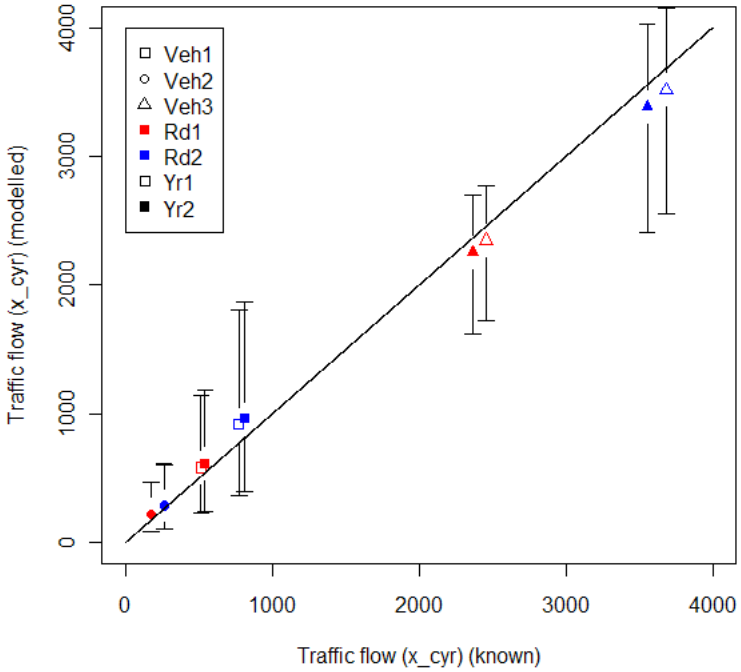


Figure 4.5: Modelled exposure (traffic flow – modelled) x_{cyr} against known test exposure (traffic flow – known) x_{cyr} for simulated data in log-Normal exposure model

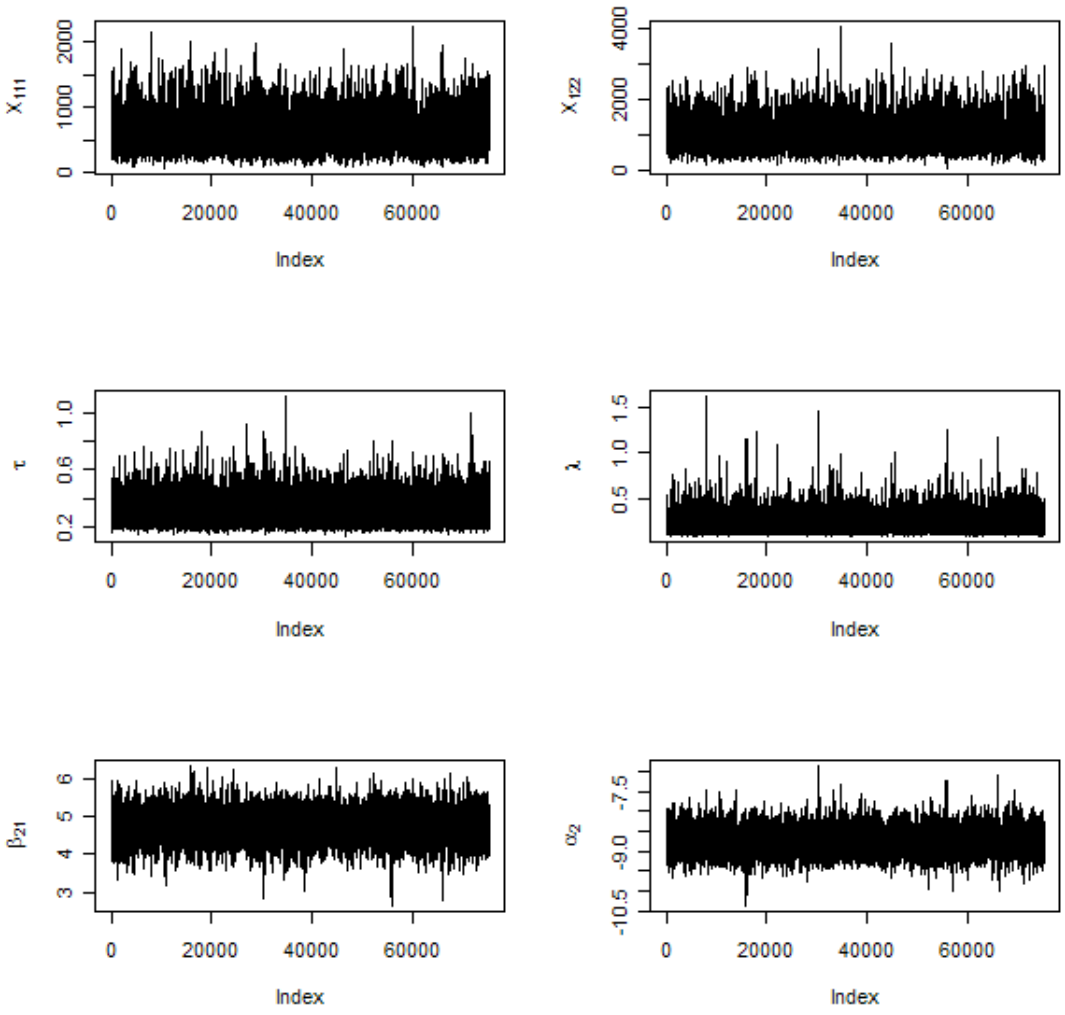


Figure 4.6: Time series of a selection of unknown parameters (iterations 25 000 to 100 000) for simulated study on log-Normal exposure model

Small road data

We have applied subsets of the known information $\boldsymbol{e}$ (induced exposure) and $\boldsymbol{z}$ (registered vehicle data) with car types 4x4s, large saloons and medium saloons ($C = 3$), years 2005 and 2006 ($Y = 2$) and road types A roads and Minor roads ($R = 2$), shown in Table 4.1 to the model. In addition, information from $\boldsymbol{x}^+$ has been used for the years and road types stated. 36% of the vehicles in the registered vehicles data were 4x4s, large saloons and medium saloons and therefore values in $\boldsymbol{x}^+$ have been factored appropriately.

The priors τ_a, τ_b, λ_a, λ_b, β_0 and α_0 are the same as above, shown in 4.21. Convergence is stable after 10 000 iterations, and a selection of parameter densities (with 10 000 iterations of burn in removed) are shown in Figure 4.7.

In order to assess this model with real data Figure 4.8 compares the mean results from the MCMC with results from the heuristic proportional fit algorithm described in Section 4.2.

Under this model, the induced exposure need not be normalised at the outset as the $\boldsymbol{\alpha}$ values vary to control for the differences across road types. The results of the model with unnormalised $\boldsymbol{e}$ (shown in the top half of Table 1.3) as an input result in very similar $\boldsymbol{x}$ (shown in Figure 4.9), $\boldsymbol{\beta}$, τ and λ estimates.

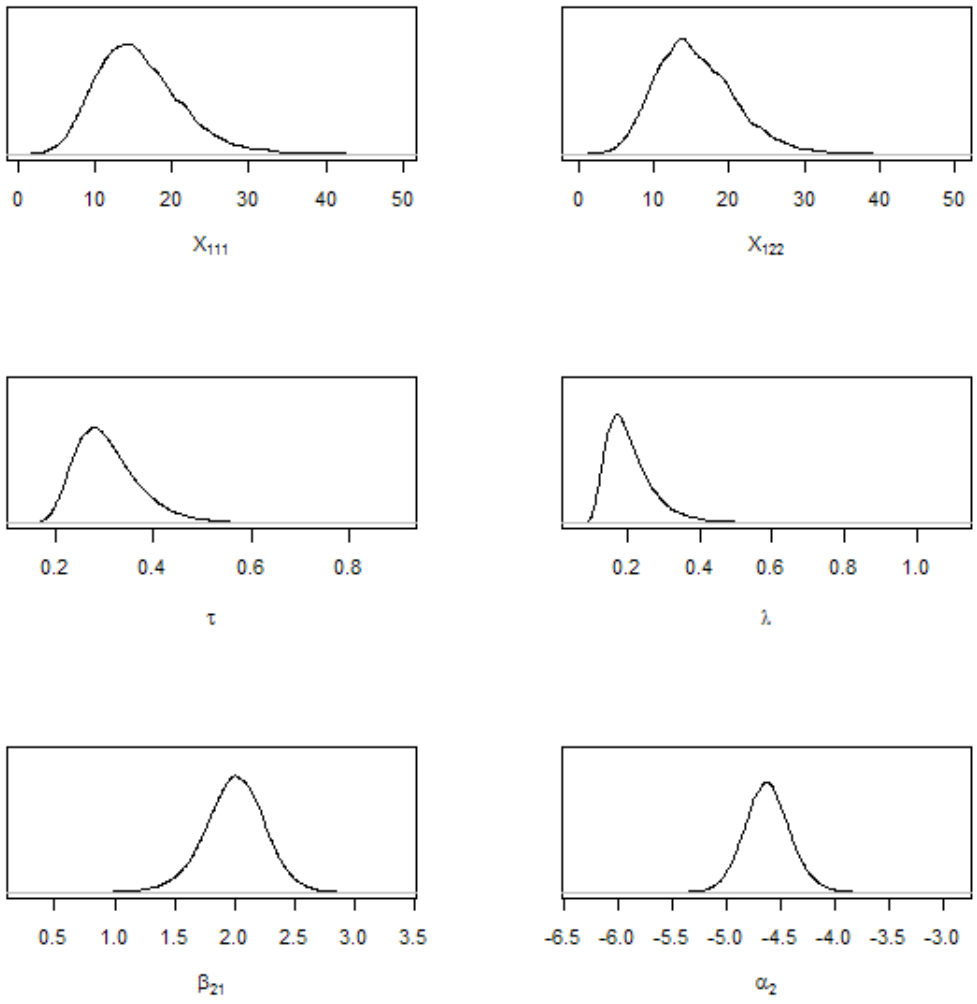


Figure 4.7: Densities of unknown parameters (iterations 10 000 to 100 000) for small dataset on log-Normal exposure model. x_{111} represents the exposure for 4x4s in 2005 on A roads, x_{122} represents the exposure for 4x4s in 2006 on Minor roads, β_{21} is based on large saloons on A roads and α_2 is the road parameter for Minor roads

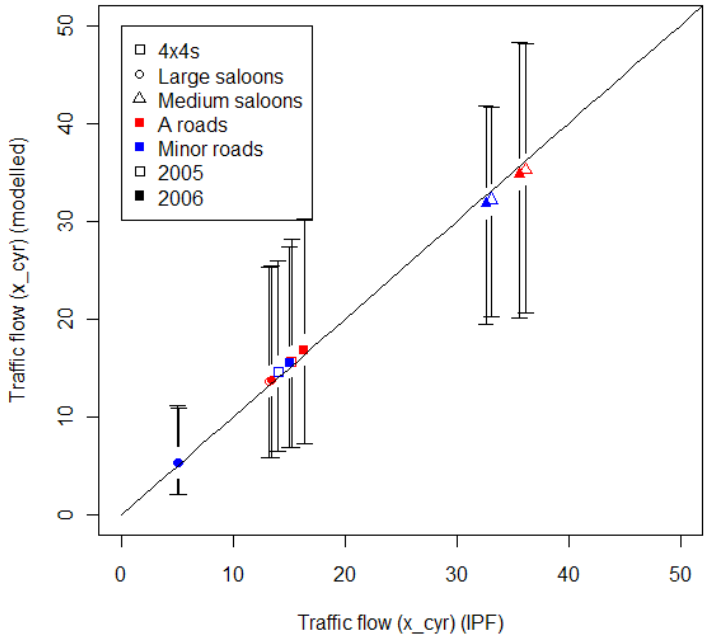


Figure 4.8: Modelled exposure (traffic flow – modelled) x_{cyr} against estimated exposure (traffic flow – IPF) x_{cyr} for small dataset on log-Normal exposure model

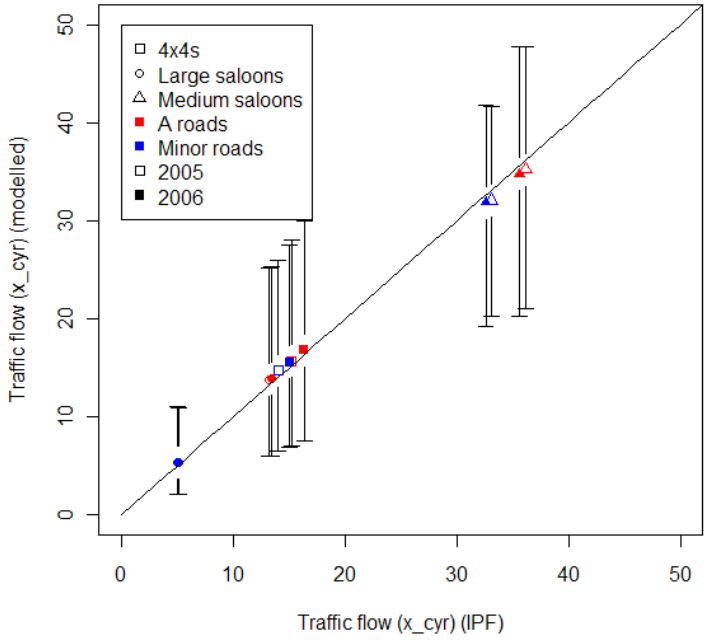


Figure 4.9: Modelled exposure (traffic flow – modelled) x_{cyr} against estimated exposure (traffic flow – IPF) x_{cyr} for small dataset with unnormalised induced exposure data as an input on log-Normal exposure model

Large road data

The large road dataset is made up of information from 12 years (1999–2010), six car types and three road types.

Four different prior distributions on τ^2 and λ^2 (shown in Table 4.13) have been used to produce four separate estimates. The first prior set (Prior 0) demonstrates the result of giving each part of the model equal weight. The specified priors (shown in Table 4.13) give approximately equal weight to the two parts of the model (shown in 4.13) with an a priori relative error of around 60%. That is, we expect that the error in $\log \boldsymbol{x}$ and $\log \boldsymbol{e}$ to be around 60% of their respective means.

The other three prior sets give less weight to the second part of the model, as we are less certain about the model representing the induced exposure data (e_{cr}). Priors on τ^2 represent the uncertainty we feel in the model assumption that the car and road does not vary by year. Priors on τ^2 are generated with the beliefs that relative errors in $\log \boldsymbol{x}$ of around 20%, 10% and 5% respectively (Prior 1, 2 and 3) exist and represent a 95% a priori probability on $\log \boldsymbol{x}$ ($\tau = 0.1$, $\tau = 0.05$, $\tau = 0.025$), i.e. diffuse, less diffuse and precise.

Priors on λ^2 must be less certain as the uncertainty we feel in this model is based on the fact that $\boldsymbol{e}$ is based on a small survey. Priors on λ^2 are generated with the beliefs that relative errors in e of around 40%, 20% and 10% respectively (Prior 1, 2 and 3) are appropriate and represent a 95% a priori probability on $\log \boldsymbol{e}$ ($\lambda = 0.2$, $\lambda = 0.1$, $\lambda = 0.05$), i.e. diffuse, less diffuse and precise.

Results discussed below are for the diffuse, weighted prior (Prior 1) unless otherwise specified, as mean results are very similar for each prior. Equivalent results for less diffuse priors (Priors 2 and 3) are shown in Appendix C.

Using a series of informal convergence monitors and a series of different starting

Table 4.13: Prior values for log-Normal exposure model on 12 year dataset

	Prior 0	Prior 1	Prior 2	Prior 3
β_0	0	0	0	0
σ_β	1 000	1 000	1 000	1 000
α_0	0	0	0	0
σ_α	1 000	1 000	1 000	1 000
τ_a	5.0	5.0	5.0	5.0
τ_b	0.5	0.06	0.01	0.003
λ_a	3	3	3	3
λ_b	0.25	0.1	0.025	0.007

values, it has been possible to see that convergence is likely to have occurred after 1 000 000 iterations. 9 000 000 further iterations have been run and every 100th iteration stored.

Figures 4.10 and 4.11 compare the results from the MCMC to the heuristic proportionally fitted estimated flows described in Section 4.2 for 4x4s. 95% errors bars show the spread of the results in the MCMC. These show that the heuristic IPF estimate gives equal weight to each part of model. Assuming a weighted certainty about each part of the model (as the diffuse prior does) results in different estimates, for example for 4x4s, the modelled values are lower on A roads and Minor roads, shown in Figure 4.11.

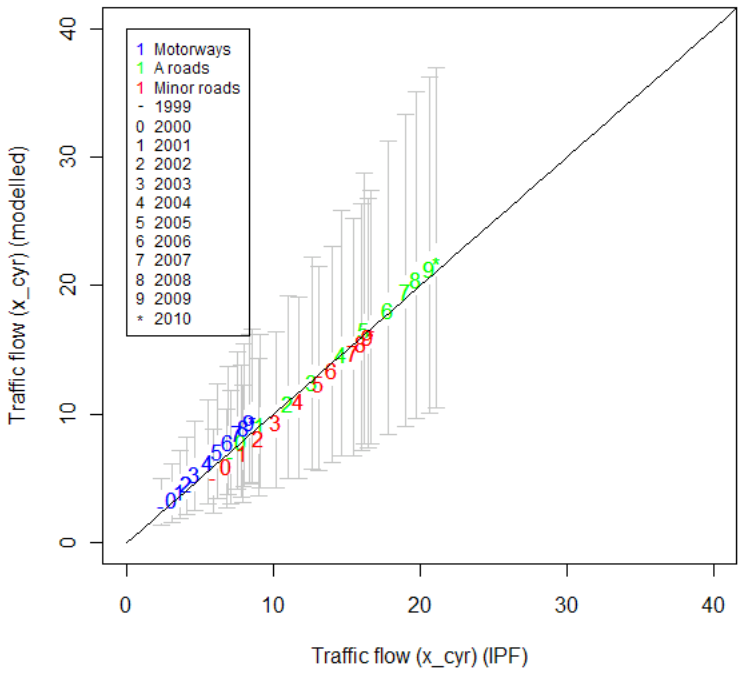


Figure 4.10: Modelled exposure (traffic flow – modelled) x_{cyr} against estimated exposure (traffic flow – IPF) for 4x4s with equally weighted prior (Prior 0)

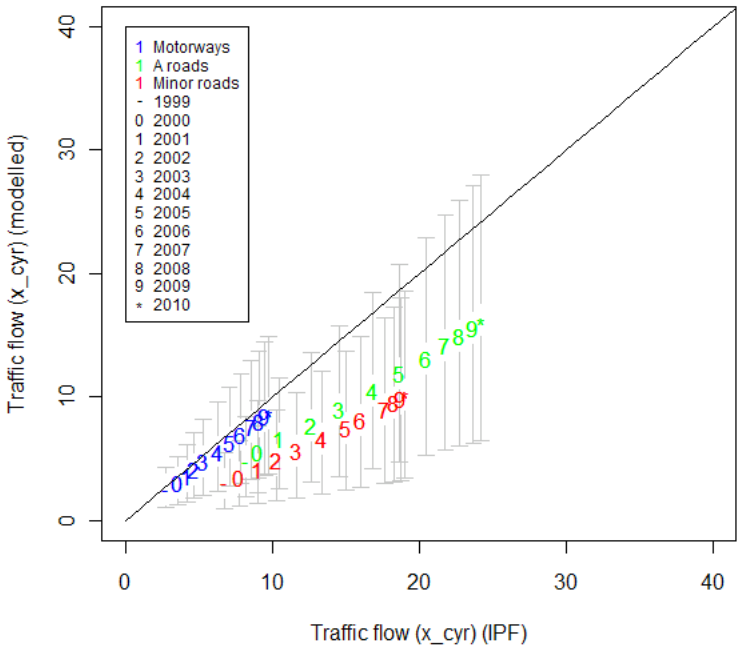


Figure 4.11: Modelled exposure (traffic flow – modelled) x_{cyr} against estimated exposure (traffic flow – IPF) for 4x4s with unequal weight and diffuse prior (Prior 1)

Figures 4.12, 4.13 and 4.14 show the modelled mean values of x_{cyr} that result from the simulation with the diffuse prior, for Motorways, A roads and Minor roads respectively. These values represent an estimate of the log of the number of billion vehicle kilometres travelled by each car type in each year on each road type, which we call disaggregated exposure data.

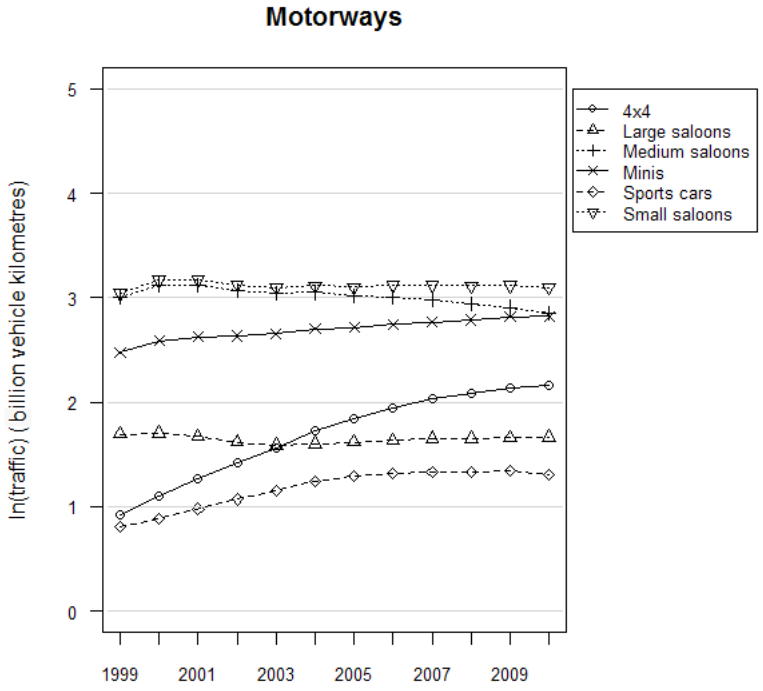


Figure 4.12: Modelled disaggregated exposure log(traffic) x_{cyr} by year and car type on Motorways from log-Normal exposure model with diffuse prior

Some clear patterns from the different datasets show up. Overall traffic levels on Motorways are considerably smaller than the other road types, Minor roads follow as 4x4s, large saloons and sports cars all contribute only a small amount of traffic on these roads, followed by A roads where a small majority of the traffic travels.

In magnitude, the car types split into two groups: 4x4s, large saloons and sports cars; and medium and small saloons and minis, with the amount of traffic due to the first group being substantially lower than that from the second group. The

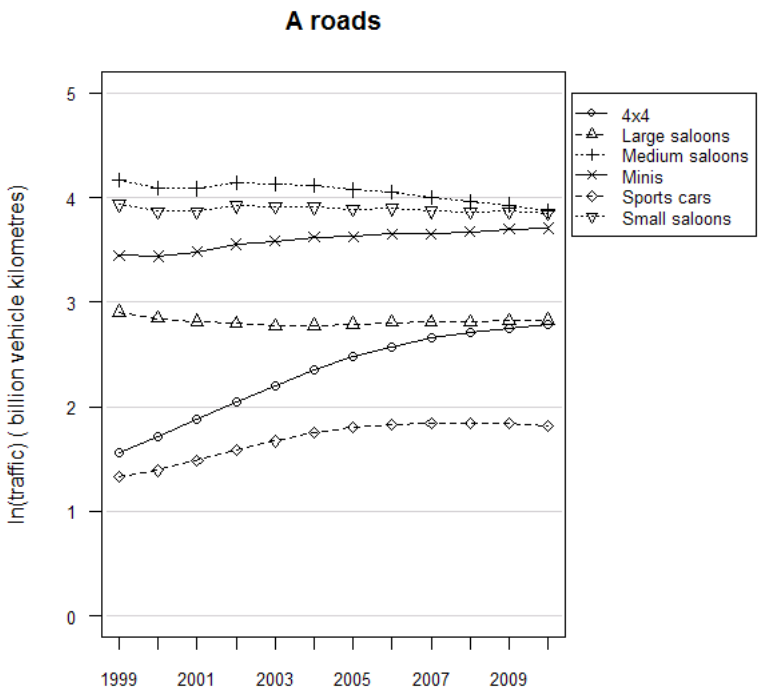


Figure 4.13: Modelled disaggregated exposure log(traffic) x_{cyr} by year and car type on A roads from log-Normal exposure model with diffuse prior

separation between the groups varies by road type with the separation biggest on Minor roads where the smaller cars (medium and small saloons and minis) are more likely to travel more than the larger cars (and sports cars). Within each of these groups the pattern varies depending on the road type with small and medium saloons making the biggest traffic contribution on Motorways, minis contributing most on A roads and small saloons (and minis by the end of the time period) being most prominent on Minor roads, although this may just be a function of the variability within the induced exposure data.

The traffic growth of 4x4s is considerable across all road types, such that, by the latest year, it is closer to the larger group than the smaller one. In addition, there are emerging patterns of traffic growth for minis and, to some extent, sports cars and some evidence of decline in the other car types.

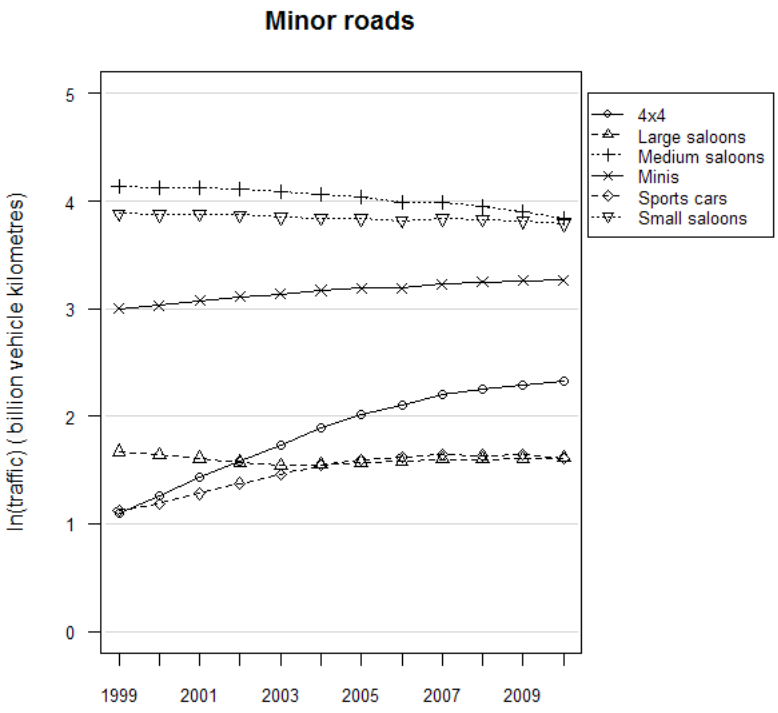


Figure 4.14: Modelled disaggregated exposure log(traffic) x_{cyr} by year and car type on Minor roads from log-Normal exposure model with diffuse prior

Figure 4.15 shows the distributions of some of the other modelled parameters after 1 000 000 iterations.

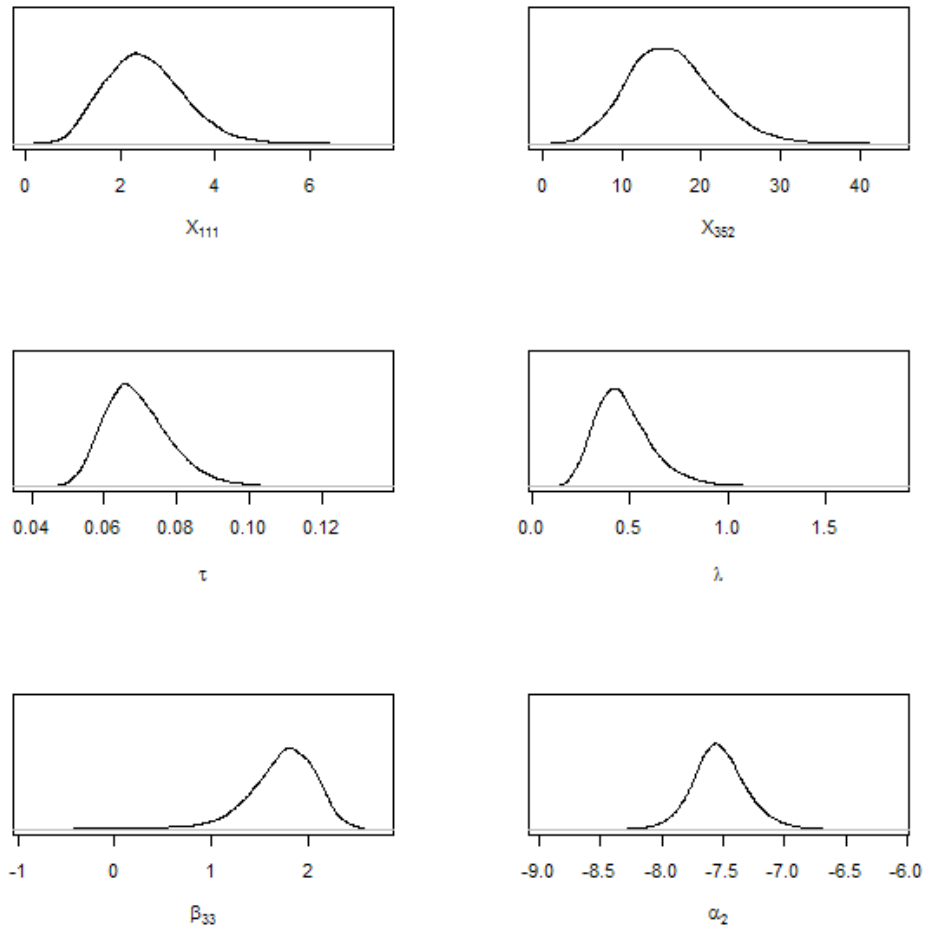


Figure 4.15: Densities of unknown parameters (every 100 iterations from 1 to 10 million) from log-Normal exposure model with diffuse prior

In Figure 4.15 x_{111} represents the exposure for 4x4s in 1999 on Motorways, x_{352} represents the exposure for 4x4s in 2010 on A roads, β_{33} is based on medium saloons on Minor roads and α_2 is the road parameter for A roads.

4.6 Estimating the distribution of the results

4.6.1 Introduction

The results from the MCMC on the large dataset (10 000 000 iterations collected every 100 over 216 parameters) are intended for use in modelling accident rates as described in Chapter 5. As we are uncertain about the exposure, the uncertainty needs to be included in the accident rate models. The following section reviews if the exposure data fit a known distribution, in particular if Multivariate Normality can be assumed. Results discussed below are for the diffuse weighted prior (Prior 1).

4.6.2 Multivariate Normality

Informal checks show that pairwise histograms of $\log \boldsymbol{x}$ produce distributions similar to a bivariate Normal (e.g. see Figure 4.16), and univariate Normality is confirmed by plotting histograms for each element of $\log \boldsymbol{x}$ however this does not confirm Multivariate Normality.

It should be noted that the covariance of all 216 parameters is singular due to the restrictions on the model: x_{+yr} was known and placed a restriction on x_{cyr} throughout the modelling process. The final 36 parameters represent the final car type. The data have been transformed to take account of this singularity, for the purposes of determining whether a standard statistical distribution can be found to approximately represent the exposure parameters:

$$X_{cyr} = \log x_{cyr} - \log x_{Cyr}$$

There are many tests referred to in the literature for testing Multivariate Normality and a few are implemented in R: the Shapiro-Wilks Normality test (Shapiro

and Wilk 1965) is extended for Multivariate Normality; a joint normal test for Multivariate Normality based on analysing Mahalanobis distances (Mahalanobis 1936); and Mardia's Test for Multinormality (Mardia 1985).

All the tests and Figure 4.17, an output from the joint normal test, suggest that Multivariate Normality should not be assumed.

4.6.3 Multivariate t-distribution

Figure 4.17 suggests that the distribution from the diffuse weighted prior may have heavier tails than a Multivariate Normal distribution allows. An obvious alternative distribution is the Multivariate t-distribution. No specific Multivariate t-distribution goodness of fit test has been found and so the following is a derived test for Multivariate t based on the Mahalonobis distance. This is a generalisation of Hotelling's T^2 statistic (Hotelling 1931).

Assume $y \sim MVN_p(0, \Sigma)$ where p is the number of dimensions and $u \sim \chi^2_k$ where k is the degrees of freedom, then

$$y\sqrt{\frac{k}{u}} = x - \mu$$

and

$$x \sim MVt_k(\mu, \Sigma)$$

Therefore

$$x - \mu \sim MVt_k(0, \Sigma)$$

Let $\Sigma = LL^T$ be the cholesky decomposition of Σ then

$$z = L^{-1}(x - \mu) \sim MVt_k(0, I)$$

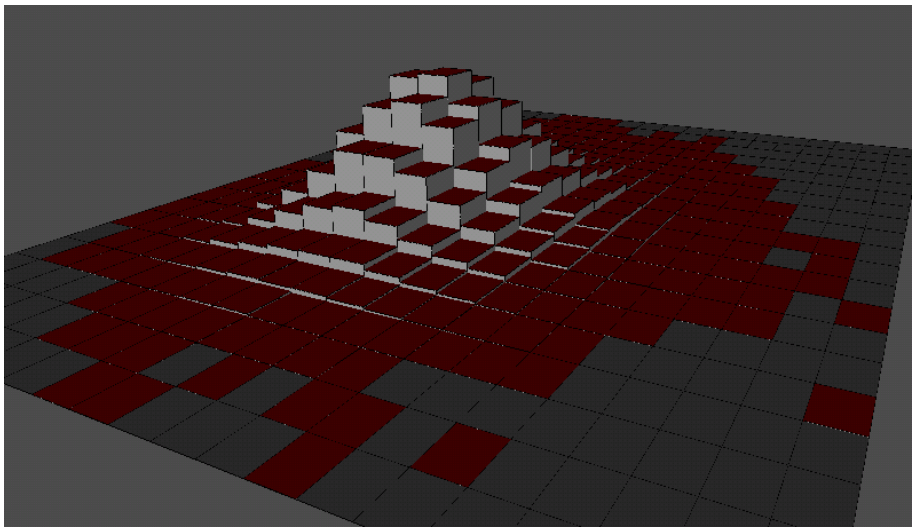


Figure 4.16: Bivariate distribution of log exposure for 4x4s on Motorways in 1999 with log exposure for sports cars on Minor roads in 2010

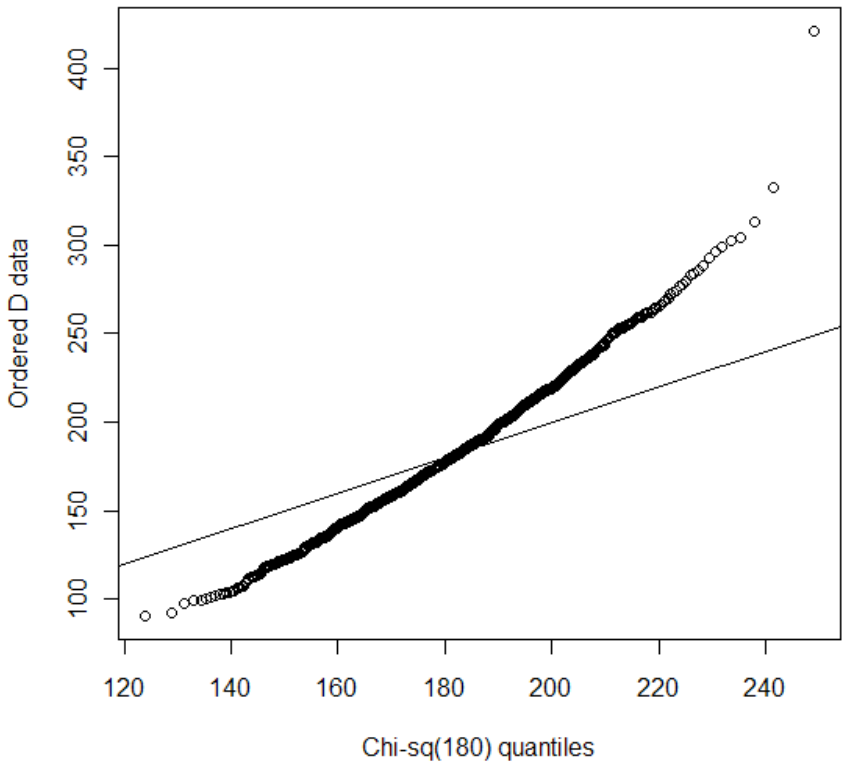


Figure 4.17: Residual plot for X_{cyr} in MVN test

and

$$\begin{aligned} z^T z/p &= \left(L^{-1}y\sqrt{\frac{k}{u}}\right)^T \left(L^{-1}y\sqrt{\frac{k}{u}}\right)/p \\ &= \frac{k}{u}y^T L^{-T}L^{-1}y/p \\ &= \frac{k}{u}y^T\Sigma^{-1}y/p \\ &= \frac{\chi_p^2/p}{\chi_k^2/k} \\ &= F(p,k) \end{aligned}$$

which implies that

$$\frac{1}{p}(x-\mu)^T\Sigma^{-1}(x-\mu) \sim F(p,k)$$

where $\Sigma \simeq \frac{k-2}{k}\hat{\Sigma}$ and $\hat{\Sigma}$ is the sample variance.

QQ-plots for a range of k are shown in Figures 4.18. These plots suggest that $k = 60$ may be a good fit.

4.6.4 Best value of k

A range of k values appear to give different QQ-plots and so we determine the best k where the correlation between the data and the theoretical quantiles is highest and the sum of the distance between the data and theoretical quantiles shown in the QQ-plots is the least. We then generate random t-distributed values to determine approximate acceptable confidence intervals and compare these intervals to the correlation and distance values.

Correlations

For a range of k, the correlation between the exact and test quantiles (in the QQ-plot) are calculated. Table 4.14 shows the correlations for a selection of different k values and a 95% confidence interval generated from randomly simulated F

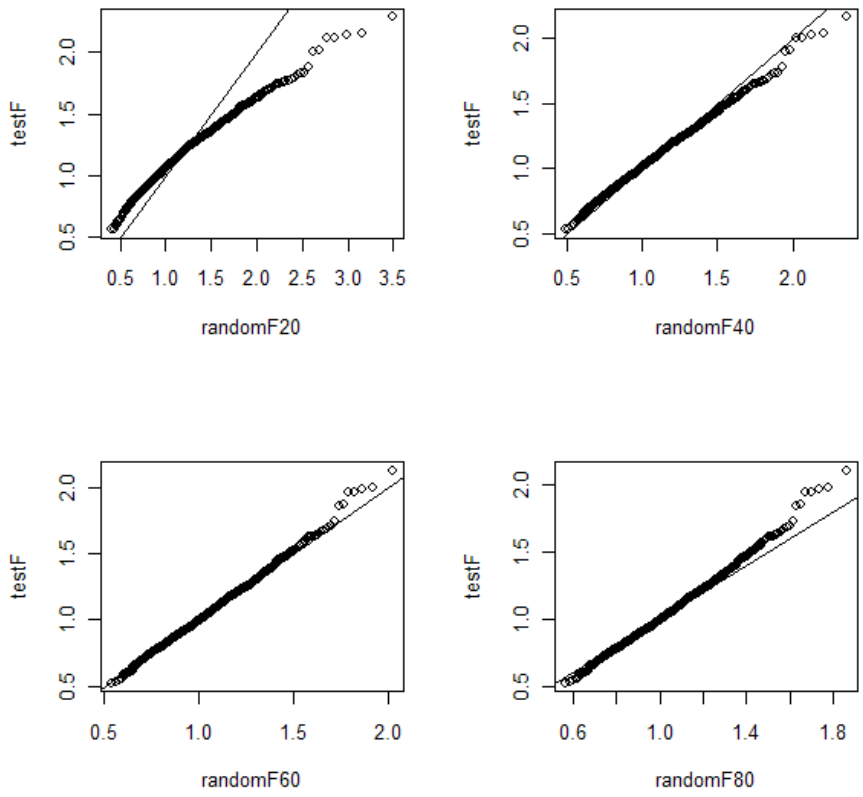


Figure 4.18: QQ-plots of exposure data on log-Normal model with diffuse prior against MVt_{20} to MVt_{80} distributions

distributed variables. The highest correlation is 0.9993 which occurs when $k = 47$, and correlations throughout the pictured range are within the simulated 95% confidence intervals.

Table 4.14: Simulated confidence intervals for correlations between exact and test quantiles in QQ-plots for testing MVt_k for exposure distribution X_{cyr} over range of k

Test correlation	Lower 2.5th percentile	Upper 2.5th percentile	k
0.9640	0.9445	0.9991	10
0.9932	0.9859	0.9994	20
0.9981	0.9908	0.9995	30
0.9992	0.9934	0.9996	40
0.9993	0.9948	0.9996	50
0.9991	0.9955	0.9996	60
0.9987	0.9958	0.9996	70
0.9983	0.9963	0.9996	80
0.9978	0.9965	0.9996	90
0.9974	0.9965	0.9996	100

Figure 4.19 shows the range of correlation values from $k = 15$ to $k = 100$ with associated confidence intervals.

Variance about the QQ-line

The correlation statistics show that values of k from 20 to 100 all give satisfactory measures of noise about the linear relationship, however it does not show that the linear relationship is of the expected form where the data and theoretical quantiles match. To determine whether any of the k values produce satisfactory QQ relationships, we have computed the sum of the absolute difference between the theoretical and data quantiles. A series of Multivariate t-distributed datasets of different k values have then been generated in order to simulate a 95% confidence interval range of acceptability.

Table 4.15 shows the results of the comparison between the data and exact quan-

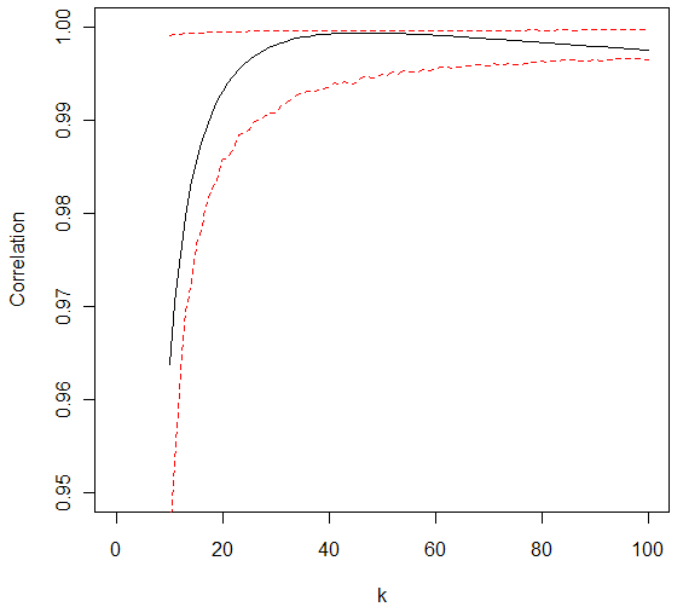


Figure 4.19: Plot of test correlations for testing MVt_k of exposure model posterior distribution over range of k – diffuse prior

tiles for a mixture of k values, and the associated acceptability bands. The degrees of freedom k with the smallest sum is $k = 57$, with a difference sum of 4.5. Comparing this to a set of randomly generated quantiles suggests that differences from $k = 45$ to $k = 70$ are approximately within a simulated 95% acceptability range.

Figure 4.20 shows the range of absolute differences between the exact and test quantiles for $k = 20, \ldots, 200$.

For the purposes of the modelling accident rates we propose a value of $k = 50$ for each prior.

4.7 Additional car age exposure information

It has been hypothesised (Broughton 2009) that the recession in 2008–2009 in Great Britain may have resulted in a change in the distribution of car ages. A reduction in the number of people buying new cars results in an increase in the

Table 4.15: Simulated confidence intervals for absolute variance from QQ-plot line for testing MVt_k of exposure model posterior distribution ($\sum_{i=1}^{10000} \mid \hat{q}_i - q_i \mid$) over range of k

Test difference	Lower 2.5th percentile	Upper 2.5th percentile	k
283.1	15.9	57.1	10
113.1	9.0	31.5	20
56.5	6.9	24.5	30
26.6	5.8	20.8	40
8.9	5.2	18.7	50
4.9	5.0	17.5	55
4.6	4.9	17.8	56
4.5	4.8	17.3	57
4.8	4.8	17.4	58
5.5	4.8	17.3	59
6.3	4.8	16.9	60
16.0	4.4	15.8	70
23.7	4.3	15.0	80
29.9	4.0	14.3	90
35.1	3.9	14.0	100

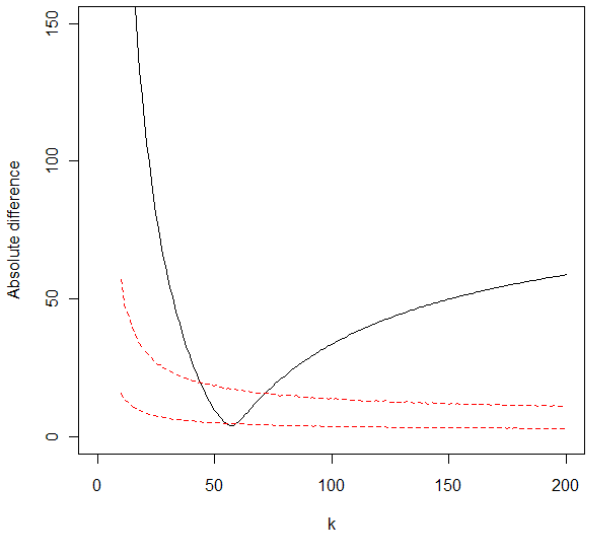


Figure 4.20: Plot of absolute variance from QQ-plot for testing MVt_k of exposure model posterior distribution over range of k – diffuse prior

overall age of the car fleet. Cars designed more recently tend to be designed to a higher safety standard and have better secondary safety features (airbags, automatic braking, electronic stability control etc.) than those designed in the past. Therefore an increase in the vehicle population's age reduces the safety of the fleet and may result in more, or more severe, accidents. In this section we describe extending the exposure modelling results for year, car type and road type to include car age to enable us to estimate the effect of possible changes of car age on accident rates.

Additional data have been received from the DVLA which contain the number of registered cars by year of first registration and car make and model by year. These data are described in Section 1.3.2.

A simple Gibbs sampler based on a log-Normal model adds some uncertainty to these data:

$$\log \boldsymbol{a} \sim N(\boldsymbol{\theta}, \delta^2)$$

generates a posterior distribution for $\log a_{cyg}$: the log of the number of registered cars by year (y), car type (c) and car age ($g = 1, \ldots, G$) given an uninformative prior on θ_{cyg}:

$$\boldsymbol{\theta} \sim N(\boldsymbol{\theta_0}, \sigma_\theta^2)$$

where $\boldsymbol{\theta_0} = 0$ and $\sigma_\theta^2 = 10$, and an a priori relative error of 25%:

$$\delta^2 \sim IG(\delta_a, \delta_b)$$

where $\delta_a = 4$ and $\delta_b = 0.05$.

The joint posterior distribution

$$
\begin{aligned}
p(\boldsymbol{\theta}, \delta^2 \mid \log \boldsymbol{a}) &\propto p(\log \boldsymbol{a} \mid \boldsymbol{\theta}, \delta^2) p(\boldsymbol{\theta}) p(\delta^2) \\
&= \Big(\frac{1}{2\pi\delta^2}\Big)^{CYG/2} \exp \sum_{cyg} \Big(- \frac{1}{2\delta^2} (\log a_{cyg} - \theta_{cyg})^2 \Big) \\
&\quad \cdot \Big(\frac{1}{2\pi\sigma_\theta^2}\Big)^{CYG/2} \exp \sum_{cyg} \Big(- \frac{1}{2\sigma_\theta^2} (\theta_{cyg} - \theta_0)^2 \Big) \\
&\quad \cdot \frac{1}{\delta^2}^{\delta_a + 1} \exp \Big(- \frac{\delta_b}{\delta^2} \Big)
\end{aligned}
$$

leads to posterior distributions for $\boldsymbol{\theta}$ and δ^2 of

$$
\begin{aligned}
\boldsymbol{\theta} &\sim N\Big(\frac{\sigma_\theta^2 \log \boldsymbol{a} + \delta^2 \theta_0}{\delta^2 + \sigma_\theta^2}, \frac{\delta^2 \sigma_\theta^2}{\delta^2 + \sigma_\theta^2} \Big) \\
\delta^2 &\sim IG\Big(\frac{CYG}{2} + \delta_a, \frac{1}{2} \sum_{cyg} \big(\log a_{cyg} - \theta_{cyg} \big)^2 + \delta_b \Big)
\end{aligned}
$$

The results of this model have been proportionally distributed over the large exposure distribution generated in Section 4.5 to give $t_{cygr} = \frac{\theta_{cyg}}{\sum_g \theta_{cyg}} \times x_{cyr}$ which is used in Chapter 5.

Due to the restrictions on data we have necessarily assumed some constant interactions, for example, the distribution of age groups across road types is constant: mathematically these restrictions are on interactions t_{cygr}, t_{cgr}, t_{ygr} and t_{gr}. Consequently the covariance matrix for t_{cygr} is singular and it is not possible to test for Multivariate Normality as we have for x_{cyr} in Section 4.6.

4.8 Discussion

In this chapter we used a Bayesian approach to model flow data disaggregated by car type. The Bayesian methodology allows us to gain information on the uncertainties in the exposure data through the priors, and results in flow estimates

with their associated uncertainty. Four models were considered which incorporated traffic flow data, registered vehicle data and induced exposure data. Initially a computationally convenient model was defined which did not use induced exposure information. This model required the assumption that the distribution of car types across different road types was the same.

A second model introduced induced exposure data which removes the need for this unjustified assumption. Exposure was modelled with a truncated Normal distribution, using two sets of priors: an imprecise and a precise set. MCMC was used to show that the precise priors produced tighter and more accurate results, and the imprecise priors led to very vague predictions. This suggests that the analysis is sensitive to the prior – there are certain aspects about which little learning is happening and we use the prior to inform these, such as the weighting applied to each dataset, but we learn from the data about other aspects. The need for a nuisance parameter, $\boldsymbol{\alpha}$, in the induced exposure part of this model was removed in the subsequent model, but large uncertainty remained.

In our final model we used the log-Normal which is a more natural way of modelling strictly positive numbers than the truncated Normal. Results based on the simulated data and two real datasets showed good convergence and matched expected results. The log-Normal structure is more convenient for deriving appropriate priors which leads to more sensible confidence intervals. The results on the full dataset clearly showed the emergence of a quickly increasing trend in 4x4 traffic.

Of the three datasets used in this paper, we are most sceptical about the reliability of the induced exposure data. We therefore apply large prior variances on these data. This leads to large variances on the final estimates. With more reliable study information a more optimistic prior would reduce the overall confidence intervals on the final estimates of exposure.

Car age data is added in at a second stage once the combination of year, road type and car type has been derived. Each two way interaction (car type and road type from the induced exposure e_{cr}, car type and year from the registered vehicle data z_{cy} and year and road type from the known traffic data x_{+yr}) is available in the initial combination of data and we therefore need to make few assumptions in deriving the three way exposure data x_{cyr}. Car age is added at a separate stage as one of the two way interactions does not exist (car age and road type) and we make the necessary assumption that cars of different ages are representatively spread across different road types.

Data estimating the two way interaction car age and road type are in theory available from the induced exposure dataset; we could determine the distribution of the age of cars on different road types based on the age of the cars not at fault in accidents. This however is flawed as we know that older cars involved in collisions are more likely to result in serious injury to its occupants than newer cars. In this case the induced exposure cannot be assumed independent of accident data and therefore cannot be used.

Each of the four datasets tell part of the story and so in combining these data the result is more flexible for analysis and more appropriate for modelling than using any of the datasets individually. These new estimates of exposure by year, road type, car type and car age can be used to monitor car traffic trends and can be applied to any car accident analysis which requires the use of accident rates. Any future analyses of different car sizes similar to that in Knowles et al. (2007); Starnes and Longthorne (2003); Broughton (2007); Keall and Newstead (2007) could use these data or the method to improve the analysis. The modelling process can be extended for additional years and could be extended to include rural and urban road types within the road types used here but is limited to the six car types due to current data constraints.

We use these results in Chapter 5 to generate accident rates.

Chapter 5

Accident rate modelling

5.1 Introduction

The main purpose of exposure data in road accident analysis is to determine whether accident numbers are high or low relative to a measure of how often a particular scenario occurs. For example, the number of sports cars involved in accidents may be small, but the total number of kilometres driven by people in sports cars is also small relative to the number of other cars. Accident numbers relative to the number of vehicle kilometres driven, that is an accident rate, will be used for analysis. The more detailed the exposure data, the fewer assumptions or limitations there are in the interpretations. Other alternatives to exposure include number of registered vehicles, length of road and population, although all of these options have limitations.

Up until now, exposure data in vehicle kilometres has been restricted to vehicle type (car, LGV, HGV, motorcycle etc.), year, month and time of day, and road type. When evaluating the relative risk of different car types, for example, being involved in an accident, other measures of exposure have been used. In Chapter

4 we have shown that it is possible to disaggregate vehicle kilometre data in its given form to encompass other variables and generate estimates for vehicle kilometres by car type and car age with associated probability distributions to reflect uncertainty. The risk of each car type in each age group to be involved in an accident can now be calculated relative to respective distances travelled.

5.1.1 Data

Accident data have been retrieved from the British personal injury road accident database STATS19, described in Section 1.2. The model detailed here is based on a subset of the STATS19 data containing killed or seriously injured (KSI) car occupant casualties involved in single vehicle accidents over a 12 year period from 1999–2010. This subset contains information on 45 394 accidents. This specific subset was chosen in order to investigate the hypotheses given in Broughton and Buckle (2007) and in particular to assess whether the rise in the number of 4x4 accidents occurring at bends from 1999–2006 was offset by the rise in 4x4 traffic.

5.1.2 The basic model

Accidents S are assumed to be Poisson distributed with some overdispersion (as discussed in Mitra and Washington 2007, among many others) and are modelled in a full factorial Poisson log-linear model with the simulated exposure data as an offset. We model car KSI accident rate (relative to an offset of exposure t_{cygr}) by variables year y (1999–2010), car type c (4x4 to sports), car age g (new to old), bend b and road class r (Motorways, A roads and Minor roads). Mathematically this is

$$S_{cygrb} \sim Po(\chi_{cygrb})$$

$$\begin{aligned}\log(\chi_{cygrb}) &= \log(t_{cygr}) + \gamma_0 + \gamma_1[c] + \gamma_2 y + \gamma_3[g] + \gamma_4[r] + \gamma_5 b + \cdots + \psi_{cygrb} \\ &= \log(\boldsymbol{t}) + \gamma_j X_{ij} + \boldsymbol{\psi}\end{aligned}$$

where t_{cygr} is traffic in billion vehicle km by car type, year, car age and road class modelled from the MCMC model in the large road data part of Section 4.7 – we make the necessary assumption that exposure at bends is the same as exposure not at bends due to data limitations. γ_1 is a vector of dimension C representing the car type coefficients, γ_2 is a constant representing the year coefficient, γ_3 is a vector of dimension G representing the car age coefficients, γ_4 is a vector of dimension R representing the road type coefficients, γ_5 is a constant representing the effect of bend and ... represent higher dimension interactions. $\gamma_1[1]$, $\gamma_3[1]$ and $\gamma_4[1]$ are constrained to 0, $\boldsymbol{\psi} \sim N(0, \sigma^2_\psi)$ is the overdispersion parameter and X_{ij} is a design matrix.

Parameters are given uninformative diffuse priors:

$$\gamma_j \sim N(0, 100)$$
$$\sigma^2_\psi \sim IG(0.01, 0.01)$$

With five variables and all associated interactions, there are too many possible models to assess the appropriateness of each model. For the purposes of this thesis we restrict ourselves to two-way interactions only and employ a search strategy to search over possible models.

5.2 Model selection

5.2.1 Search strategy

Here we base the model choice on a Bayesian marginal likelihood method where marginal likelihoods are computed using the Laplace approximation (see Section 3.9.3). In Section 4.7 we discussed that it was not possible to estimate t_{cygr} using a standard distribution due to internal conditional independences. We therefore

account for the variability in the exposure data by randomly selecting a number of draws from the posterior distribution of t_{cygr} and run the model selection algorithm on each draw rather than including it in the marginal likelihood representation.

The search strategy is a repeated 10 step process:

1. Pick a draw from the posterior distribution of t_{cygr}.
2. Compute the marginal likelihood of the main effects model.
3. Add each two-way interaction individually to the main effects model and estimate the marginal likelihood.
4. Calculate the model probabilities for the set of models in the previous step.
5. Accept the model with the highest model probability.
6. Register any models with a model probability which is higher than 75% of the highest model probability.
7. Add each remaining two-way interaction individually to the best model from step 5.
8. Repeat from step 4 until a full model is reached.
9. Repeat the process for any models registered in step 6.
10. Repeat the whole process (drawing a different t_{cygr}) n times.

Model probabilities are calculated (weighted appropriately to take into account registered models) for all the models tested and the best model or models are chosen.

5.2.2 Marginal likelihood approximation

Estimation of the marginal likelihood uses the Laplace approximation (described in Section 3.9.3) on the posterior distribution derived below to find the highest marginal likelihood. Here we use i as an index representing the combined indices

cygrb for simplicity.

$$
\begin{aligned}
p(S \mid \gamma, \psi) &= \int \prod_{ij} p(S_i \mid \chi_i) p(\gamma_j \mid \gamma_0, \sigma_\gamma^2) p(\psi_i \mid \phi_0, \sigma_\psi^2) p(\sigma_\psi \mid \psi_a, \psi_b)\, \mathrm{d}\gamma\ \mathrm{d}\psi\ \mathrm{d}\sigma_\psi \\
&= \int \prod_{ij} \frac{\exp\left[-t_i e^{\gamma_j X_{ij} + \psi_i}\right] \left[t_i e^{\gamma_j X_{ij} + \psi_i}\right]^{S_i}}{S_i!} \\
&\qquad \cdot \frac{1}{(2\pi)^{1/2}\sigma_\gamma} \exp\left[-\frac{1}{2}\left(\frac{\gamma_j - \gamma_0}{\sigma_\gamma}\right)^2\right] \\
&\qquad \cdot \frac{1}{(2\pi)^{1/2}\sigma_\psi} \exp\left[-\frac{1}{2}\left(\frac{\psi_i - \psi_0}{\sigma_\psi}\right)^2\right] \\
&\qquad \cdot \frac{\psi_b^{\psi_a}}{\Gamma(\psi_a)} \sigma_\psi^{-(\psi_a+1)} \exp\left(-\frac{\psi_b}{\sigma_\psi}\right)\ \mathrm{d}\gamma\ \mathrm{d}\psi\ \mathrm{d}\sigma_\psi \\
&= \int \prod_{ij} \frac{\exp\left[-t_i e^{\gamma_j X_{ij} + \psi_i}\right] \left[t_i e^{\gamma_j X_{ij} + \psi_i}\right]^{S_i}}{S_i!} \\
&\qquad \cdot \frac{1}{(2\pi)^{1/2}\sigma_\gamma} \exp\left[-\frac{1}{2}\left(\frac{\gamma_j - \gamma_0}{\sigma_\gamma}\right)^2\right] \\
&\qquad \cdot \frac{1}{(2\pi)^{1/2} e^{\omega}} \exp\left[-\frac{1}{2}\left(\frac{\psi_i - \psi_0}{e^{\omega}}\right)^2\right] \cdot \frac{\psi_b^{\psi_a}}{\Gamma(\psi_a)} e^{-\omega(\psi_a+2)} \\
&\qquad \cdot \exp\left(-\frac{\psi_b}{e^{\omega}}\right)\ \mathrm{d}\gamma\ \mathrm{d}\psi\ \mathrm{d}\omega
\end{aligned}
$$

where $\omega = \log(\sigma_\psi)$ for simplification. This integral can be approximated by Laplace approximation.

$$
p(S) = Q \int g(\gamma, \psi, \omega) = Q \int g(\boldsymbol{\delta}) \quad \simeq Q g(\hat{\boldsymbol{\delta}}) | - h''(\hat{\boldsymbol{\delta}})|^{-1/2} (2\pi)^{d/2}
$$

where $d = \dim(\boldsymbol{\delta}) = \dim(\gamma, \psi, \omega) = J + 1 + I$ and

$$
Q = \prod_i \left[\frac{t_i^{S_i}}{S_i!}\right] (2\pi)^{-J/2} \sigma_\gamma^{-J} (2\pi)^{-I/2} \frac{\psi_b^{\psi_a}}{\Gamma(\psi_a)}
$$

$$g(\hat{\boldsymbol{\delta}}) = \exp\left(\sum_i \sum_j \left[\left\{-t_i e^{\hat{\gamma}_j X_{ij} + \hat{\psi}_i}\right\} + \left\{\hat{\gamma}_j X_{ij} + \hat{\psi}_i\right\} S_i\right] - \sum_j \left[\frac{1}{2}\left(\frac{\hat{\gamma}_j - \gamma_0}{\sigma_\gamma}\right)^2\right]\right.$$
$$\left.\sum_i \left[\hat{\omega} - \frac{1}{2}\left(\frac{\hat{\psi}_i - \psi_0}{e^{\hat{\omega}}}\right)^2\right] - \hat{\omega}(\psi_a + 2) - \psi_b e^{-\hat{\omega}}\right)$$

$$h(\boldsymbol{\delta}) = \log(g(\boldsymbol{\delta}))$$

$$\hat{\boldsymbol{\delta}} = \operatorname{argmax} g(\boldsymbol{\delta})$$

5.2.3 Results

Table 5.1 contains a subset of the results from the model selection strategy (the complete table is shown in Appendix D as Table D.1) for one exposure run. The added two-way interaction is shown in column 2 followed by the estimated marginal likelihood and associated model probability in columns 3 and 4. Column 5 shows the BIC value for comparison. Results based on a selection of further exposure runs resulted in selection of the same models with similar weights. There was one model registered[1] – model 43 had a model probability very close to model 45. We ran further models adding in interactions from model 43 which resulted in the same choice of high probability models.

The two models which have been selected (based on the marginal likelihoods) are models 43 (ME+cr+yg+rb+cg+gr+cb^2) and 45 (ME+cr+yg+rb+cg+gr+yb). These models include 97% of the model probability and the results from each model will be averaged to give an overall severity rate for each category. The BIC would have selected the more complex model 54: $ME + cr + rb + cg + yg + gr + yb + cb + cy + yr$.

[1]Models are registered if they have a model probability which is higher than 75% of the highest model probability within one step of the model selection strategy (see Section 5.2.1)

[2]ME is the main effects model

Table 5.1: Subset of marginal likelihoods and model probabilities for accident rate model selection

	Model	Marginal likelihood	Model choice	BIC
ME + cr + rb + cg + yg				
36	+cy	40738.8	0%	12353.8
37	+cb	40746.5	0%	12357.2
38	+yr	40744.6	0%	12362.6
39	+yb	40748.2	0%	12361.3
40	+gr	40749.1	1%	11989.6
41	+gb	40743.4	0%	12378.8
ME + cr + rb + cg + yg + gr				
42	+cy	40747.1	0%	11978.6
43	+cb	40753.7	48%	11982.1
44	+yr	40746.6	0%	11983.1
45	+yb	40753.7	49%	11986.1
46	+gb	40744.0	0%	12015.6
ME + cr + rb + cg + yg + gr + yb				
47	+cy	40743.5	0%	11974.7
48	+cb	40750.5	2%	11979.5
49	+yr	40743.6	0%	11980.9
50	+gb	40740.8	0%	12011.0

5.3 Modelling accident rate with mean exposure

Initially a model is fitted with a fixed exposure variable: the mean values from the exposure model shown in Figures 4.12, 4.13 and 4.14. Posterior distributions have been generated in winBUGs (Lunn et al. 2000) based on 100 000 iterations, with a thinning frequency of 10 and the first 100 samples were removed for burn-in. Informal convergence monitors suggest convergence has been reached.

As it is a large model, the parameters are shown in Table D.2 in Appendix D. A large σ^2_ϕ value shows, as is common with accident data, that there is overdispersion in the data.

Figure 5.1 shows the model averaged accident rates for each factor split into its

categories. We replicate all accidents within each factor and so each accident is represented five times in Figure 5.1 and the following similar graphs.

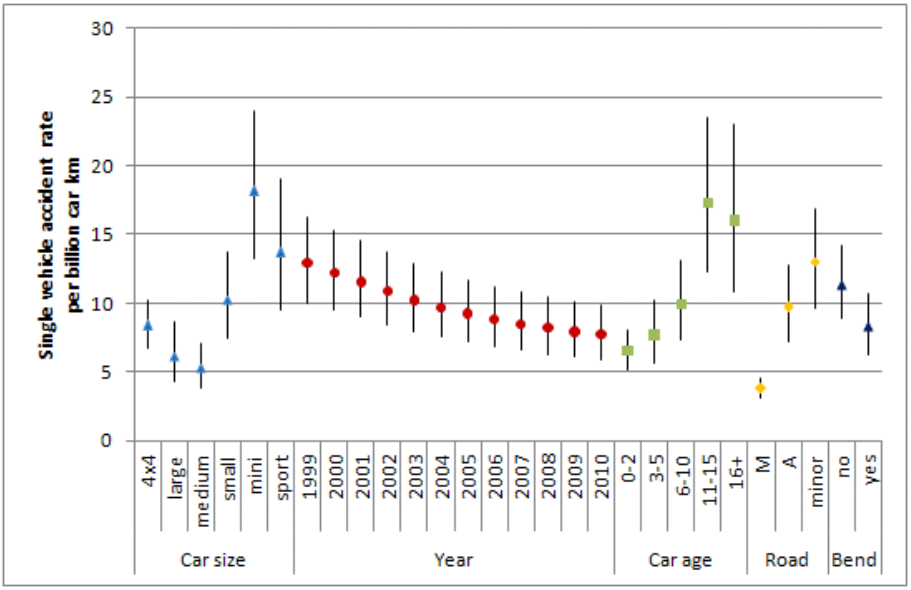


Figure 5.1: Medians and 95% posterior intervals of model averaged accident rates for single vehicle car accidents by main factors in fixed exposure model

The accident rate for killed or seriously injured (KSI) car occupants is considerably higher on A roads and higher still on Minor roads relative to Motorways. This is consistent with findings in the annual report of road casualties (Department for Transport 2011) for all vehicles and severities, albeit more extreme than here: we observe a relative KSI accident rate for A roads and Minor roads of around 2.5 and 3.5 times bigger than Motorways, whereas for all vehicles the relative accident rate for A roads and Minor roads are around 6.0 and 7.5 bigger respectively (Department for Transport 2011). KSI accident rates for car occupants are highest for minis and superminis with a rate twice as high as 4x4s and people carriers. This is consistent with findings in Knowles et al. (2007). Older cars have higher accident rates, most likely due to continual improvements in protection within cars perhaps making a fatal or serious accident in an old car into a less severe accident in a new car.

As this model does not take into account the uncertainty in the exposure, the variability around the estimates is considerably underestimated.

5.4 Modelling accident rate with variable exposure

5.4.1 Generating exposure and rates in simulation

In order to propagate uncertainty from the exposure model into the accident rate model we must incorporate the exposure posterior distribution into the accident rate MCMC using the basic model form described in Section 5.1.2. In a standard MCMC structure with both the uncertain exposure distribution and the accident data as inputs, the MCMC would update everything and we would learn about the exposure data from the accident data (and vice versa). It is the different relationships of these two concepts, exposure and accidents, across the different road user groups which is of interest here so we made the prior decision that accident data will not be used to update uncertainty about the exposure distribution.

There are two main computational methods for incorporating uncertain exposure into the MCMC without allowing this distribution to be updated by information from the accident rate model. Firstly, in WinBUGs, it is possible to generate a new draw from the exposure distribution in each iteration separately from updating the accident rate model and therefore not allow information from the accident data to update the exposure distribution. This requires the exposure distribution to be represented as a functional form.

In Section 4.6 we showed that the traffic exposure measure over year, road and car type could be represented by a Multivariate t-distribution with 50 degrees of freedom. Due to restraints on the data, it was not possible to represent the

larger exposure measure t_{cygr} using a simple statistical distribution. However it is possible to generate a MVt-distribution over x_{cyr} followed by a Normal distribution over a_{cyg} and calculate the combination described in Section 4.7 within each MCMC iteration.

Alternatively and for reasons of computing time here, we select a scenario approach where each MCMC accident rate model is run over a series of different exposure runs (t_{cygr}) generated in the exposure modelling process. We combine the resulting accident rate models as described in Section 5.4.2 using a model averaging process. The MCMC was run for 5 000 iterations over each of 300 randomly selected exposure runs from the posterior distribution of the exposure modelling process. For each model over each exposure run a thinning frequency of 10 was applied and the first 100 iterations were removed for burn-in.

5.4.2 Model results

We have used WinBUGs (Lunn et al. 2000) to run a total of 150 000 iterations over a selection of 300 randomly selected exposure runs for each high probability model (models 43 and 45) derived in Section 5.2.3. We give parameters uninformative priors and model average over the two models.

Once again, we show the median and standard deviations of coefficients for the two high probability accident rate models in Appendix D, in Table D.3. The standard deviation values for each coefficient are considerably larger than those in Table D.2 as these models now contain variability from the exposure measure. Figure 5.2 shows the accident rates and respective posterior intervals for each factor. The pattern of the means is similar to that shown in Figure 5.1 except that the median rates for 4x4s and large saloons have swapped over.

We assess the model fit with a posterior predictive check: across the 2 160 combi-

nations of the five factors, the observed accident numbers fall within the predicted confidence bands in 92.8%, however these confidence bands are large. For example, for all single vehicle car accidents occurring in 2010, the model estimates that the accident rate (per billion car km) was 7.7 with a 95% posterior range from 1.6 to 25.5. A range this large is not that useful for future predictions, so in Section 5.4.3 we investigate reducing the variability.

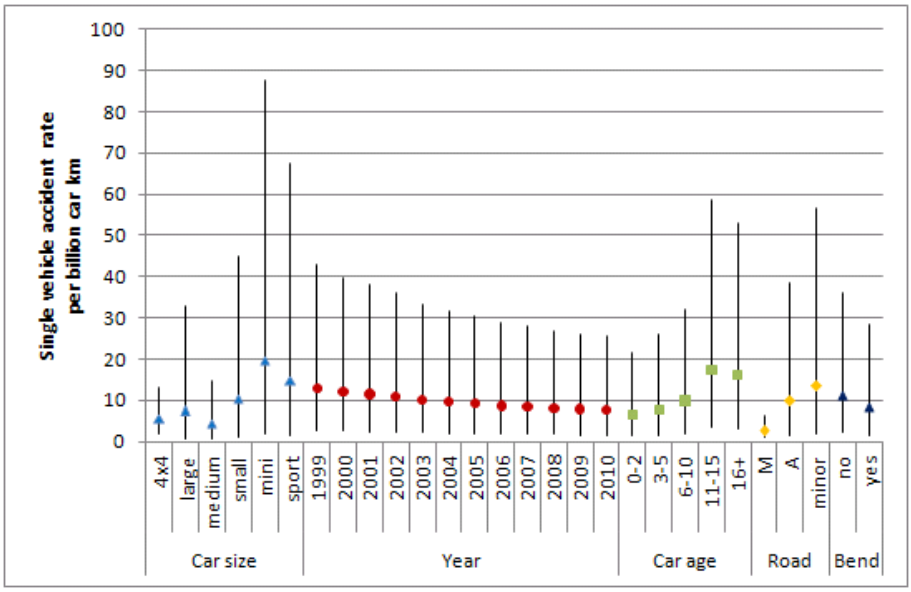


Figure 5.2: Medians and 95% posterior intervals on accident rates for single vehicle car accidents by main factors with variability in exposure

5.4.3 Reducing variability

In Section 4.5 we estimate the exposure for the large dataset on four different sets of priors, getting progressively more certain about the data. Up until now we have used the results from the most diffuse prior. In order to increase the use of any future predictions using these models, we have reduced the estimated uncertainty in the exposure measure, thereby reducing the variability in the posterior estimates of the accident rates. Figure 5.3 shows the equivalent graph to Figure 5.2 over this reduced uncertainty. In general the posterior intervals have reduced by around 40%. The estimate for large saloons has increased, and it is thought that this is

due to the small number of large saloon accidents meaning that only small changes are required to result in big variation in the accident rate.

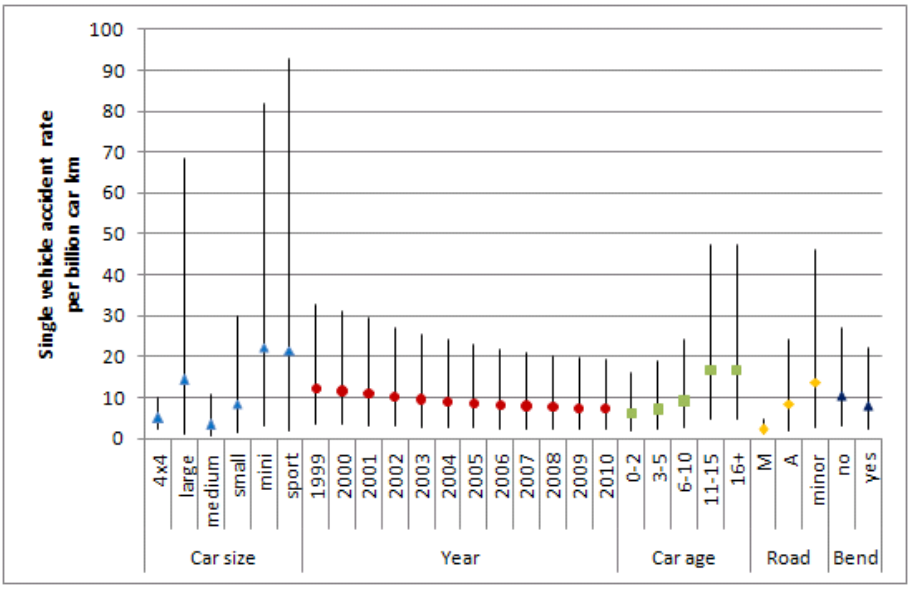


Figure 5.3: Medians and 95% posterior intervals on accident rates for single vehicle car accidents by main factors with reduced variability in exposure measure

5.5 Introducing economic factors

It has been suggested (Broughton 2009) that the economic position of the country affects the number of severe road accidents that occur. In particular, in times of recession there appears to be drops in the number of fatal accidents which are not possible to explain with other factors. Here we use a measure of economy (GDP per capita, described in Section 1.3.5) to replace the linear factor year to investigate whether economy is a better predictor of fatal and serious accident rate for single vehicle accidents.

Table 5.2 contains the marginal likelihood values for each high probability model, firstly with the factor year (from Table 5.1) and secondly from the model with economy replacing year. These latter values are considerably lower than the year

model and show that introducing this measure of economy does not improve the model.

Table 5.2: Comparison of marginal likelihood values for high probability accident rate models including factors year or economy

	Model	Year model	Economy model
ME + cr + rb + cg + yg + gr			
43	+cb	40753.7	40689.4
45	+yb	40753.7	40692.7

This may be due to the pattern of road accident numbers in relation to the shape of the economy measure over time – in general the number of road accidents decreases over time, with a sharper drop in fatal accidents in times of recession. The economy measure increases in general, with drops in times of recession. Alternatively the increased drop in fatal accidents may not have been big enough to have affected the trend in the number of serious and fatal accidents. We expect that the trend in the number of fatal accidents may be affected by the economy, and this is investigated in Section 6.4.

5.6 Discussion

The accident rate model discussed in this chapter uses the modelled exposure data to derive accident rates across many more categories than have been possible previously. Due to the further disaggregation of exposure data, accident rates for car type, car age, road type and year concurrently can now be estimated, and include associated uncertainty propagated through from the exposure measure derived in Chapter 4. These rates have been modelled as a Poisson model with exposure as an offset.

The model selection process, using Laplace approximation to estimate the marginal likelihoods, showed that two models with interactions between car type and road

(cr), year and car age (yg), road and bend (rb), car type and car age (cg), car age and road (gr) and either car type and bend (cb) or year and bend (yb) were the best fit from the set of models with two way interactions. The results of the two models have been averaged to give overall accident rates for each combination of the five main effect factors car type c, year y, car age g, road type r and bend b. The median accident rates suggest that these particular groups have higher accident rates than others:

- minis and superminis, and sports cars compared to other car types;
- older cars;
- accidents occurring on Minor roads relative to other road types; and
- accidents in earlier years.

Interpreting the effect of bend on accident rate is not possible at this stage as there exists no data on the prevalence of bends, or indeed the definition of a bend, on roads and therefore we have made the simplifying assumption that drivers spend as many miles driving around bends on each road type – an assumption which we know not to be true and which shows up in the interaction bend and road type discussed below.

In Chapter 2 we discuss the relationship between accident numbers and rates, and how it is important to consider both when evaluating results by categorising points into high, medium and low priority as shown in Figure 2.1. High rates and counts are high priority for interventions. Figure 5.4 presents the median results based on the main factors car type, car age and road type. Obviously these results have some variability associated with them, but these are large and make it difficult to observe the main picture. Figure D.1 in Appendix D shows these for the factor car type.

The car type data (red points) suggest that minis should be targeted – they have both a high accident count and rate relative to the other car types. Small saloons

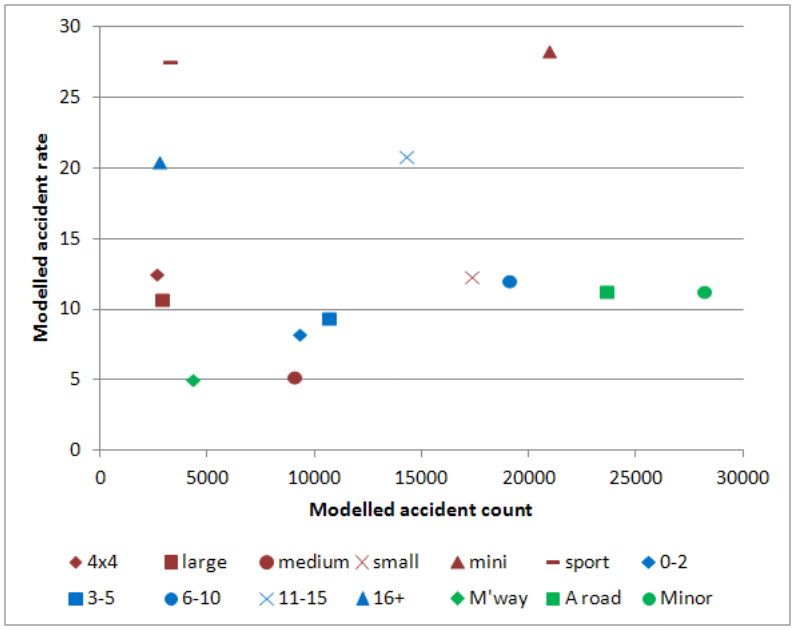


Figure 5.4: Relationship between modelled posterior median accident rates and accident propensities for factors car type, car age and road type

and sports cars fall into the medium category with high propensity, low risk and low propensity, high risk respectively. As discussed in Section 2.1 these are difficult to deal with effectively and economically. The remaining three categories are relatively low risk and low propensity. For age groups (blue points) cars aged 11-15 years are relatively high risk and high propensity making these the most obvious category to target, although older cars also fall into the high risk group. By road type (green points), A and Minor roads appear to be equivalent in terms of risk, but accidents occur more often on Minor roads.

There were six interactions in both models, of which five were consistent. The model averaged accident rates for these five interactions (in the fixed exposure model, for demonstration purposes) are shown in Figure 5.5.

The first figure contains the interaction between car type and age and shows that, for all but sports cars, the accident rate is higher for older cars than younger cars. For sports cars the pattern is slightly different with cars in the oldest category

having a lower accident rate than all other ages. This category is likely to contain 'classic' cars and these will generally be used in different driving conditions and with different driving styles than 'old' cars. We have not included this sort of variability in the exposure data.

The graph at the top right of Figure 5.5 shows the accident rates for the interaction between car type and road type. For all car types Motorway accident rates are considerably lower than other roads. In general the accident rate on Minor roads is higher than that on A roads (except for medium saloons where it is approximately the same); for some car types (minis and large saloons in particular) the accident rate on Minor roads is considerably higher.

In the second row of graphs, the interaction between road and bend shows that the numbers[3] of accidents at bends and not at bends are much closer on Minor roads than A roads and Motorways, and this is due to the fact that Minor roads have more bends.

The accident rates for the interaction of car age with road type is shown in the fourth graph in Figure 5.5. It shows that as cars get older the difference in accident rate across the road types increases. This may be due to a limitation in the exposure data which does not allow for different road use by older and younger cars; due to data limitations we had to assume that the use of all aged cars was the same across different road types. It is likely that younger cars are more likely to be driven on larger roads (i.e. Motorways and A roads) than Minor roads, and this may account for some of the differences seen in this graph.

The final graph shows the changing spread of accident rates by age group across the 12 year period. In the early years the difference in accident rate between the younger cars (up to 10 years old) and older cars (11 or more years old) was considerably bigger than in 2010. This pattern is discussed in Section 1.3.2 and

[3]Exposure at bends is unknown so we have assumed equal exposure at and not at bends

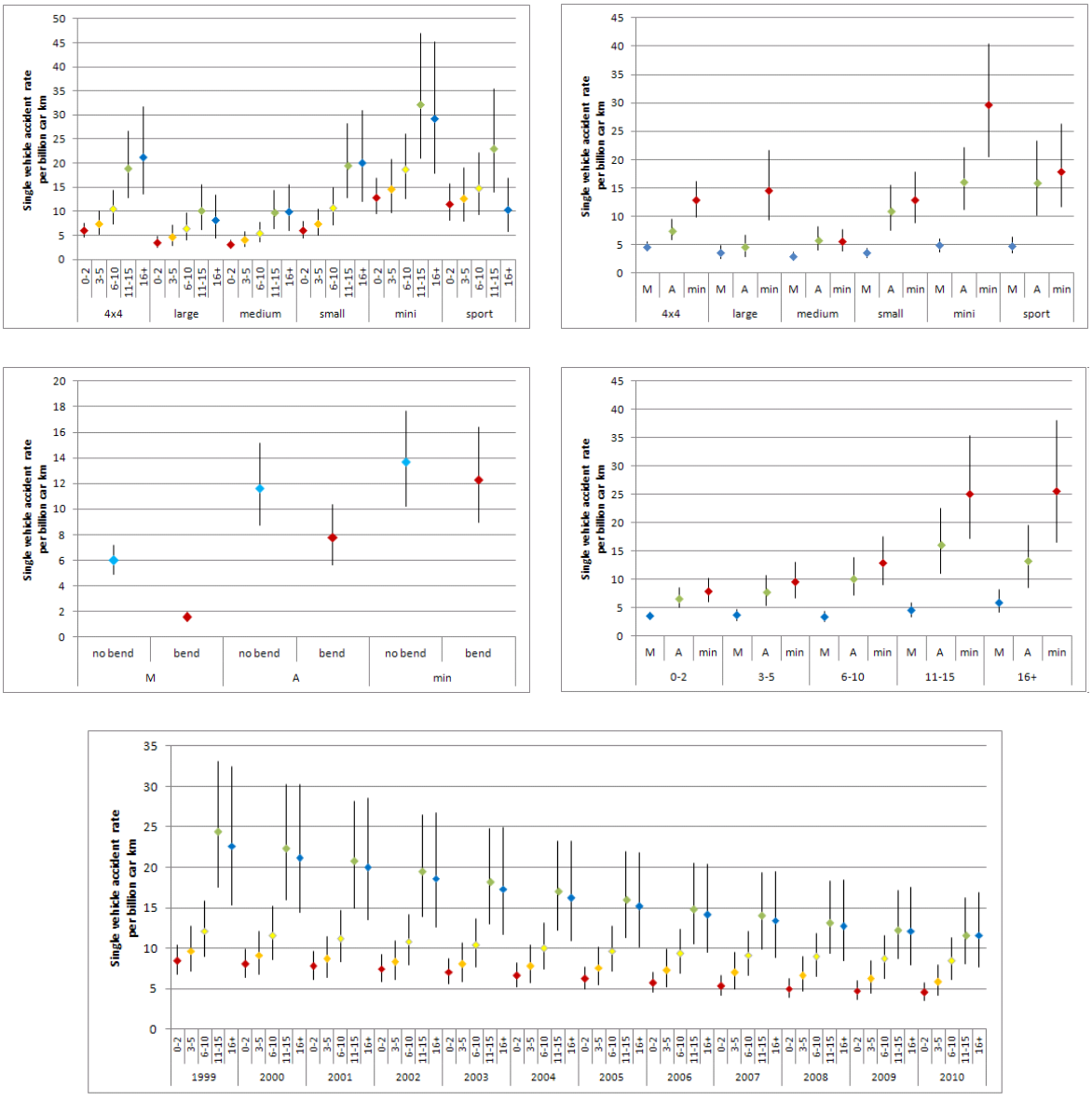


Figure 5.5: Medians and 95% posterior intervals on model averaged accident rates on fixed exposure for model interactions car type and age, car type and road type, road type and bend, car age and road type, and car age and year

is thought to be due to improving secondary safety features.

We encompass three levels of variability in the exposure data in three separate models; firstly we assume that the exposure measure is fixed and has no associated variability, next we assume a large amount of variability based on a diffuse prior specification for the exposure model in Section 4.5 and finally a reduced variability based on a tight prior for the exposure model in Section 4.5. In general the mean accident rates show the same pattern however the variability in the posterior intervals vary by model. The fixed exposure model results in approximately symmetric intervals of size around 60% of the median. The diffuse (Prior 1) exposure model variability is much larger as it encompasses variability in the exposure measure. These intervals are positively skewed and are of size around 300–400% of the median. The precise (Prior 3) model also results in positively skewed intervals which are in general of length around 200–300% of the median. Both the diffuse and precise models lead to very uncertain results and this shows how difficult it is to generate useful estimates and future predictions whilst incorporating all known uncertainty.

We attempted to improve the model by replacing the constant factor year with a measure of economy. Successive recessions have coincided with significant drops in the number of fatal accidents and we investigated whether we could show a correlation between fatal and serious accidents and a measure of economy. In this case the model containing economy as a factor rather than year was not as good a model and this suggests that the rapid decrease in the number of fatal accidents over the recession is not sufficient to have affected the overall trend in fatal and serious accidents enough to notice an effect. An investigation into severity rates in Chapter 6 will develop this investigation.

Chapter 6

Accident severity modelling

6.1 Introduction

The third modelling task is defined to assess the severity of an accident once it has occurred. Here we assess whether a number of factors can be used to predict whether the most severely injured car occupant in a single vehicle accident is seriously or fatally injured. We use the single vehicle accident dataset (based on STATS19) where at least one car occupant has been fatally or seriously injured. In addition to the factors used in Chapter 5 (car type, year, car age, road and bend) we also know whether the most severely injured occupant was fatally or seriously injured and we introduce a variable which defines whether the car overturned: Broughton and Buckle (2007) suggest that the number of cars overturning was an accident type of concern in the mid to late 1990s.

The model is derived to estimate the proportion of all accidents where a car occupant is killed or seriously injured (S) which are fatal (F). A simple binomial model uses the logit link to regress on each of the main effects and appropriate

interactions X for each of I categories[1] i. The unknown coefficients are represented as ϕ_j where $j = 1, \ldots, J$ and are estimated using Bayesian parameter estimation techniques.

$$F_i \sim Bin(p_i, S_i)$$
$$\text{logit}(p_i) = \sum_{j=1}^{J} \phi_j X_{ij} \qquad i = 1, \ldots I$$

6.2 Model selection

A model selection process, similar to that described in Section 5.2.1, is used to assess which model is best. We assume that each main effect (car type c, year y, road r, car age g, overturn o and bend b) is necessary and use the Laplace approximation (described in Section 3.9.3) on the posterior distribution derived below to find the highest marginal likelihood.

The joint posterior distribution is a combination of the likelihood:

$$\begin{aligned} p(F \mid \boldsymbol{\phi}) &\propto \prod_i \left(\frac{\exp(\sum_j \phi_j X_{ij})}{1 + \exp(\sum_j \phi_j X_{ij})} \right)^{F_i} \left(\frac{1}{1 + \exp(\sum_j \phi_j X_{ij})} \right)^{S_i - F_i} \\ &= \prod_i \exp \left\{ \sum_j \phi_j X_{ij} F_i - F_i \log \left(1 + \exp \left[\sum_j \phi_j X_{ij} \right] \right) \right. \\ &\qquad \left. - (S_i - F_i) \log \left(1 + \exp \left[\sum_j \phi_j X_{ij} \right] \right) \right\} \\ &= \prod_i \exp \left\{ \sum_j \phi_j X_{ij} F_i - S_i \log \left(1 + \exp \left[\sum_j \phi_j X_{ij} \right] \right) \right\} \end{aligned}$$

[1] a category is a combination of factors where there is a least one accident

and the prior: $\phi_j \sim N(\phi_0, \sigma_\phi^2)$

$$p(\phi) = Q \prod_j \frac{1}{\sigma_\phi} \exp\left\{ -\frac{1}{2} \left(\frac{\phi_j - \phi_0}{\sigma_\phi} \right)^2 \right\}$$

where $Q = \prod_j \frac{1}{\sqrt{2\pi}} = (2\pi)^{-d/2}$, $d = \max(j)$ giving

$$\begin{aligned} p(\boldsymbol{\phi} \mid F) \propto \int \prod_{ij} \exp\left\{ \sum_j \phi_j X_{ij} F_i - S_i \log\left(1 + \exp[\sum_j \phi_j X_{ij}] \right) \right\} \\ .Q \frac{1}{\sigma_\phi} \prod_j \exp\left\{ -\frac{1}{2} \left(\frac{\phi_j - \phi_0}{\sigma_\phi} \right)^2 \right\} \mathrm{d}\phi \\ \propto Q \frac{1}{\sigma_\phi} \prod_j \int \exp\left\{ \sum_i \left(\phi_j X_{ij} F_i - S_i \log\left(1 + \exp[\phi_j X_{ij}]\right) \right) \right. \\ \left. - \frac{1}{2} \left(\frac{\phi_j - \phi_0}{\sigma_\phi} \right)^2 \right\} \mathrm{d}\phi \end{aligned}$$

As described above, the integral is approximated using Laplace approximation.

With six main effects the number of interactions is large and therefore we simplify the problem by assuming only two-way interactions are of interest and applying a search strategy. The results of the search strategy are shown in Table 6.1 and are referred to by model number in the following strategy:

1. Estimate the marginal likelihood for the main effects model (1).
2. Add each two-way interaction individually to the main effects model and estimate the marginal likelihood (2–16).
3. Accept the model with the highest marginal likelihood from the previous step (11).
4. Add each remaining two-way interaction individually to the best model from the previous step (17–30).
5. Repeat from step 3 until a full model is reached.

Table 6.1 contains a subset of the results from the model selection strategy (the complete table is shown in Appendix E as Table E.1). The added two-way interaction is shown in column 2 followed by the estimated marginal likelihood and associated model probability in columns 3 and 4. Column 5 shows the BIC value for comparison.

The three models which have been selected are models 11 (ME+gr), 16 (ME+ob) and 30 ($ME + gr + ob$). These models include 92% of model probability and the results from each model will be averaged to give an overall severity rate for each category. The results over BIC are slightly different, although the model with the lowest BIC (model 16) is included in the model average.

6.3 Model results

Table 6.2 contains the results of the three models run in winBUGs with uninformative priors (the prior for each coefficient was $\phi_j \sim N(0, 100)$). Each model was run over 100 000 iterations with a thinning frequency of 10 and the first 1 000 iterations were removed for burn-in. The models showed good convergence and mixing, and did not appear to be sensitive to changes to prior specification as long as they remained vague.

From each model a mean severity rate (fatal/fatal and serious accidents) was estimated for each category. These were averaged to produce an overall estimated severity rate for each category using the model probabilities (factored up to 100%) shown in Table 6.1. Figure 6.1 shows the modelled severity rate for one category across each model and the model average, showing that the model average result has smaller confidence intervals than the individual model predictions. We have used posterior predictive checks (discussed in Section 3.10) to evaluate the model fit. Of the 3 647 categories where there was at least one accident, posterior pre-

Table 6.1: Subset of marginal likelihoods and model probabilities for accident severity model selection

	Model	Marginal likelihood	Model choice	BIC
1	ME	-17032.3	0%	9236.9
2	+cy	-17050.4	0%	9688.3
3	+cg	-17069.1	0%	9433.2
4	+cr	-17051.6	0%	9358.1
5	+co	-17034.9	0%	9304.5
6	+cb	-17041.5	0%	9317.3
7	+yg	-17046.1	0%	9612.3
8	+yr	-17038.2	0%	9464.1
9	+yo	-17032.6	0%	9386.8
10	+yb	-17032.4	0%	9386.4
11	+gr	-17025.6	16%	9331.5
12	+go	-17035.7	0%	9296.9
13	+gb	-17040.1	0%	9305.3
14	+ro	-17033.2	0%	9276.0
15	+rb	-17030.5	0%	9270.7
16	+ob	-17026.0	10%	9253.4
	ME + gr			
17	+cy	-17048.6	0%	9782.9
18	+cg	-17067.2	0%	9529.3
19	+cr	-17049.8	0%	9452.8
20	+co	-17033.1	0%	9362.3
21	+cb	-17039.6	0%	9374.6
22	+yg	-17044.3	0%	9706.8
23	+yr	-17036.3	0%	9558.7
24	+yo	-17030.8	0%	9407.1
25	+yb	-17030.6	0%	9406.6
26	+go	-17033.9	0%	9391.2
27	+gb	-17038.2	0%	9369.7
28	+ro	-17031.4	0%	9352.1
29	+rb	-17028.7	1%	9347.0
30	+ob	-17024.2	65%	9335.6

Table 6.2: Mean and standard deviation of coefficients for high probability accident severity models

		Model 11		Model 16		Model 30	
		Mean	SD	Mean	SD	Mean	SD
Constant		-2.15	0.12	-2.06	0.09	-2.18	0.12
Vehicle	4x4	-	-	-	-	-	-
	Large	0.34	0.09	0.35	0.09	0.35	0.09
	Medium	0.21	0.07	0.21	0.07	0.21	0.07
	Minis	-0.12	0.07	-0.12	0.07	-0.12	0.07
	Sports	0.34	0.09	0.34	0.09	0.34	0.09
	Small	0.03	0.07	0.03	0.07	0.03	0.07
Year		0.02	0.00	0.02	0.00	0.02	0.00
Car age	0–2	-	-	-	-	-	-
	3–5	0.02	0.15	-0.01	0.05	0.00	0.15
	6–10	0.24	0.14	0.00	0.04	0.23	0.14
	11–15	0.17	0.16	0.01	0.05	0.16	0.16
	16+	0.38	0.23	0.08	0.07	0.38	0.23
Road	M	-	-	-	-	-	-
	A	-0.01	0.12	-0.11	0.05	-0.02	0.12
	Minor	-0.09	0.12	-0.27	0.06	-0.09	0.12
Overturn		0.13	0.03	0.22	0.04	0.22	0.04
Bend		0.21	0.03	0.28	0.04	0.28	0.04
Age &	0–2 & M	-	-			-	-
road	0–2 & A	-	-			-	-
	0–2 & Min	-	-			-	-
	3–5 & M	-	-			-	-
	3–5 & A	-0.03	0.17			-0.02	0.16
	3–5 & Minor	-0.02	0.17			-0.01	0.17
	6–10 & M	-	-			-	-
	6–10 & A	-0.23	0.15			-0.22	0.15
	6–10 & Minor	-0.31	0.15			-0.30	0.15
	11–15 & M	-	-			-	-
	11–15 & A	-0.07	0.17			-0.06	0.17
	11–15 & Minor	-0.29	0.18			-0.28	0.17
	16+ & M	-	-			-	-
	16+ & A	-0.36	0.26			-0.37	0.26
	16+ & Minor	-0.32	0.25			-0.32	0.25
Overturn &	Bend			-0.19	0.06	-0.19	0.06
Deviance		9122.37	6.77	9119.58	5.64	9113.10	6.87

dictive checks show 98.4% of the observed data categories fall within the expected confidence intervals.

6.4 Introducing economic factors

The pattern of severity rates over 1990–2010 is discussed in Section 1.1 and saw a general increase from 1990 to 2006, due to the rise in the number of fatal accidents of certain accident types such as single vehicle accidents, accidents at bends and accidents where the vehicle overturned. From 2007 the number of fatal accidents declined rapidly and with it the severity rate (proportion of fatal accidents over all fatal and serious accidents) and it is hypothesised that this change may have been affected by the economy. Here we replace the main effect year with the nominal GDP per capita (described in Section 1.3.5; in £k) to evaluate whether a measure of the economy can be used as a predictor in the model.

Table 6.3 shows that the marginal likelihoods for the models with interactions between overturn and bend and age and road are slightly lower than the same model with year replaced by economy (-17026 compared to -17019 for example). Table E.2 contains the parameters of the model along with their associated standard deviations – these have changed very little from Table 6.2. As above we have used the posterior predictive method to assess the model fit. Across the 3 647 categories, the observed severity rates fall within the predicted confidence bands in 98.3% of cases. This is the first statistical suggestion that the economy has an effect on the severity of accidents, however neither coefficient is large, for year or economy, but these are significant and comparable in size to other significant main effects. The relationship between economy and accident severity cannot be confirmed as a causal relationship without considerably more work in this area.

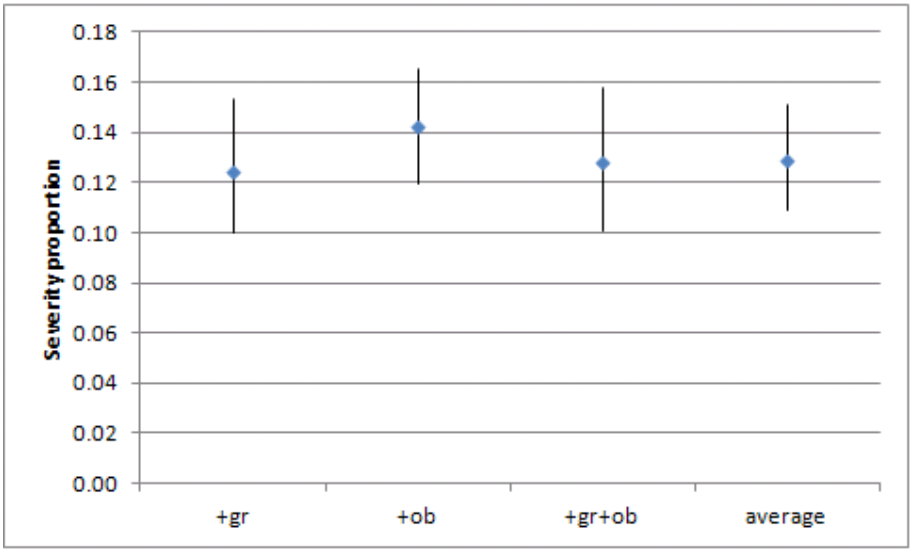


Figure 6.1: Means and 95% posterior intervals of modelled severity proportion for each high probability severity model for new 4x4s on Motorways in 1999 which did not overturn and not at a bend

Table 6.3: Comparison of marginal likelihoods and model probabilities for accident severity models including factors year or economy

	Year model		Economy model	
Model	Marginal likelihood	Model choice	Marginal likelihood	Model choice
ME	-17032.3	0%	-17025.1	0%
+gr	-17025.6	16%	-17018.7	16%
+ob	-17026.0	10%	-17019.1	10%
+ob+gr	-17024.2	65%	-17017.3	65%

6.5 Discussion

Reported in this chapter are two sets of models which appear to accurately model accident severities across different car types, road environment (road type and bend), accident type (overturn) and year. The models assume all main effects are valid and a model selection process suggests that three models with two way interactions age and road ($+gr$), over and bend ($+ob$) and the two interactions combined, can be used to adequately model the severity rates. The results of these three models have been averaged to give overall model results and these averages suggest that particular groups result in less severe injuries when involved in accidents. In particular, lower severity rates are predicted for:

- minis and superminis, 4x4s and people carriers, and small saloons compared to the other car types;
- accidents occurring on Minor roads compared to the other road types;
- accidents not occurring at a bend;
- accidents where the car did not overturn;
- accidents in earlier years; and
- accidents which involved cars aged under 16 years.

Accidents of interest where the severity rate was notably higher included:

- large saloons and sports car; and
- cars older than 15 years.

Many of these results are expected, such as accidents where the car overturns are generally more severe, and some results can be explained by other influences, for example, young and less experienced drivers may own older cars and accidents on Minor roads are likely on average to be at lower speeds. Several less expected results deserve some discussion however.

There appear to be significant differences between expected severities in a single vehicle accident dependent on car type, with the very large and small cars resulting in less severe injuries than others. 4x4s and people carriers tend to have stiffer exteriors than smaller cars and should therefore protect the car occupants from higher severities than less stiff cars. That small saloons and minis have lower severity rates is interesting – this may be due to the types of driver and driving styles (for example speed choice) that these cars are associated with, although there is no significant interaction of road type and car type, so the common belief that smaller cars are used for shorter journeys on Minor roads is not explicitly demonstrated here.

The age of a car also affects the severity rate once involved in an accident. Secondary safety developments, as discussed in Section 1.3.2, mean that within each accident year newer cars are generally better equipped to protect occupants than older cars in that same year. In addition, you would expect to see that new cars in more recent years are better at protecting their occupants than new cars earlier in the period, and this would have been expected to show in a significant interaction between age and year which was not seen. The distinction between severity rates across the different car age groups is not as obvious as expected, with the average modelled severity rates of all cars under 16 year being similar, and only those older than 15 years having substantially higher modelled severity rates.

Similar to the concepts discussed in Section 5.6, common, high severity accidents are of most concern. Figure 6.2 presents severity proportion (number of fatal accidents over fatal and serious all accidents) plotted with actual number of accidents for each factor. Each colour represents a different factor, and contains all accidents in the dataset.

For car type (red points) the highest severity occurs in large saloons and sports cars, however these are rare accidents. Accidents occurring in 4x4s are similarly

rare but of a much lower severity. Some further investigation is required to determine why there is such a big difference in the severity rates for 4x4s and large saloons. Medium and small saloons and minis fall in severity rate as the total number of accidents increases, with the least severe, and most common, accidents occurring for occupants in minis and superminis. This is likely to be influenced by type of drivers and driving style of these small cars – perhaps mainly used for short trips.

Car age (blue points) has been split into five groups and the choice of groups influences the pattern in the graph. For example if we have chosen to combine cars aged 0-2 years and 3-5 years into one group this would be one of the biggest groups of accidents and of reasonable concern as the severity rate is not substantially lower than any age group. The newer cars do not appear to provide their occupants with much better protection when involved in an accident.

Overturn (yellow) and bend (purple) accidents occur very much as expected, if a car overturns, then it is more likely to result in a fatal injury than a car which does not overturn, and accidents at bends are more likely to result in a fatality than those not at bends. Obviously these factors are related: an accident at a bend is more likely to result in an overturned vehicle, and this interaction is evidenced in the model selection process which selects models with the two-way interaction between overturn and bend.

Different road types (green points) show that accidents occurring on Motorways result in the highest severities but are relatively rare, accidents occurring on A roads and Minor roads are more common and result in lower severity injuries with Minor road accidents resulting in the lowest severity. These patterns are related to average speeds on these different road types – accidents occurring at fast speeds are more likely to result in more severe injuries due to kinetic energy dispersion (European Commission 2012).

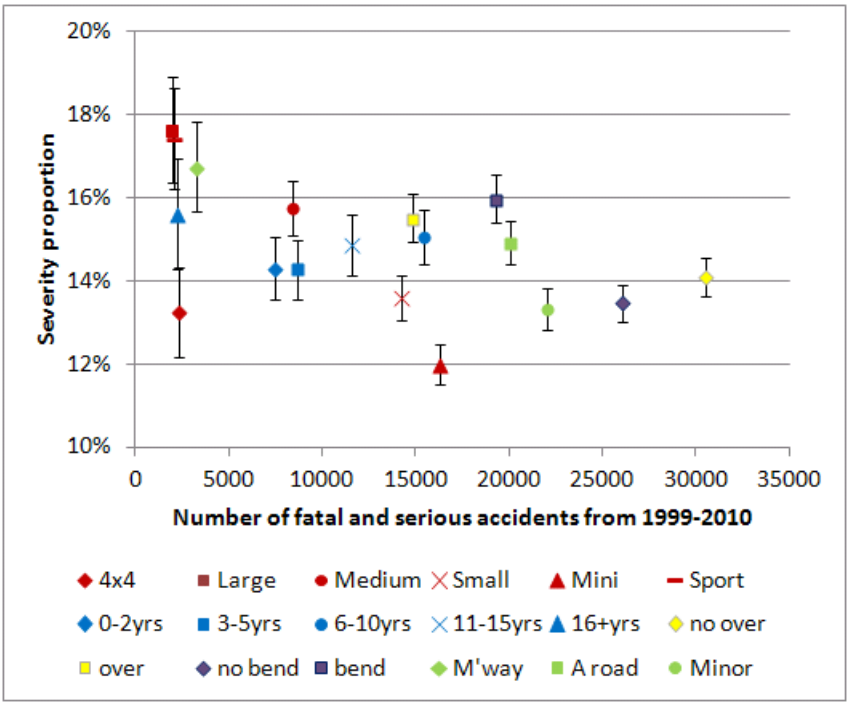


Figure 6.2: Relationship between modelled accident severity and actual accident numbers for factors car type, car age, overturn, bend and road type

The second interaction selected by the model selection procedure was the interaction between age and road type. Figure 6.3 shows this interaction in terms of severity rate and number, similarly to Figure 6.2, except each accident is only represented once in this graph. We see that the pattern of severity across car ages varies dramatically by road type. Older cars (6 years or older) involved in accidents on Motorways have a much higher severity rate than younger cars. On other roads the severity rates are similar for most age groups, excluding 11-15 year old cars which are substantially more likely to be fatal on A roads than on Minor roads. This suggests that the added protection that new cars have to reduce occupant injury starts to have a substantial effect once involved in a high speed collision.

Perhaps of most concern, and those which should be prioritised for intervention where possible, are the clump of data points of medium severity and medium

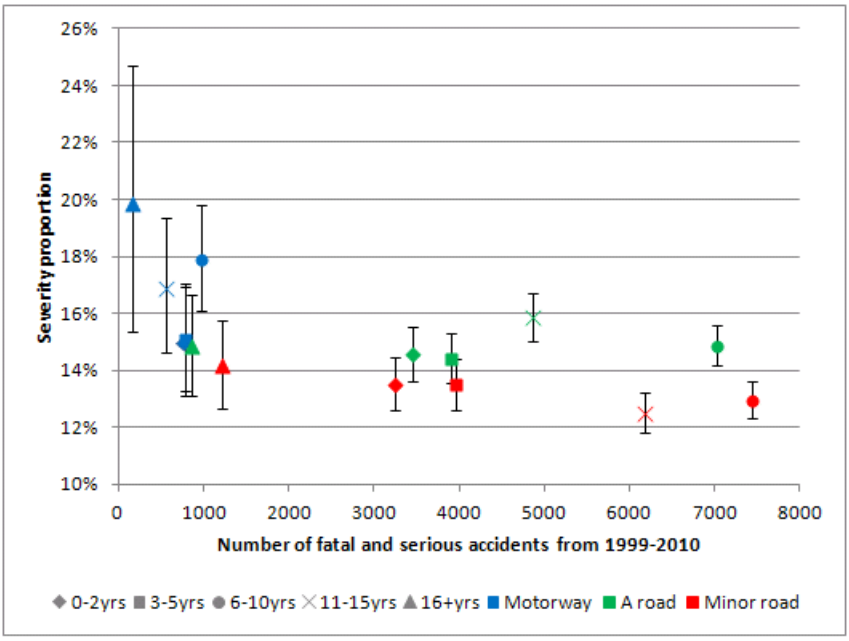


Figure 6.3: Relationship between modelled accident severity and actual accident numbers for age road interaction

– high counts in Figure 6.2: accidents occurring at bends, those where the car overturns, those on A roads and accidents in cars aged 6-10 years. Improvements in the severity rates of older cars should continue as old cars are replaced naturally by newer cars with better safety features – older cars in 2010 have more safety features than old cars in 1999 and this should continue to improve over time.

The second model in this section introduces a measure of economy as an explanatory factor in place of year. This model appears to be better in terms of its marginal likelihood compared to the associated models with year and it *suggests* that the economy may be affecting the severity of accidents. If this is the case, and this is very difficult to prove definitively, then it is likely that it is an indirect effect; for example Lloyd et al. (2013) show that there has been a reduction in young drivers and average speeds over the period where the economy was declining which is likely to have an effect on casualty severities.

In general we would expect the severity rate to have decreased over time (or over an increase in economy) due to improvements in safety on the roads, however the models suggest the opposite. The pattern in the observed severity rates appears to increase to a peak around 2005–2006, and is followed by a drop in severity rates. This suggests that the linear factor year may not have been the most appropriate measure to include and explains why the measure of economy which is not monotonically increasing improves the model. This suggests that there are other factors that are not used in these models that are important and would improve future models. This may include factors relating to weather conditions as periods of very cold weather result in fewer accidents as fewer people drive and they drive more carefully, or an average speed measure, although it would be difficult to capture sufficient variability in any speed measure. It is due to this uncertainty that we cannot explicitly say that the economy is having a direct impact on casualty severities.

Chapter 7

Predicting forward using Graphical Models

7.1 Motivation

One of the advantages of a Bayesian methodology is that it is possible to combine several models together. This is useful in this situation where the exposure, accident rate and accident severity models lead on from one another. Uncertainty in each model can be propagated through to subsequent models in a joint model framework. Here we use a graphical model to represent the combination of the models, and use this structure to predict future accident rates and numbers.

7.2 Introduction to Graphical Models

A graphical model is a graph which characterizes a probabilistic model. The graph consists of nodes (variables) and a set of edges which connect the nodes and represent dependencies between variables. These edges can be directed (arrows)

or undirected (edges).

A Bayesian Network is a Directed Acyclic Graphical model (DAG), where directed means that each edge is an arrow with a start node (parent) and an end node (child) and acyclic defines a graph with no cycles[1].

Nodes joined together with edges are dependent variables and nodes that are not joined together are conditionally independent given the values of their parents. The combination of these directed edges represent the set of conditional distributions which form the joint probability distribution.

The joint probability model can be read from the graph. The graph in Figure 7.1 represents the joint probability density function:

$$p(A, B, C, D) = p(A)p(B)p(C|A)p(D|A, B) \tag{7.1}$$

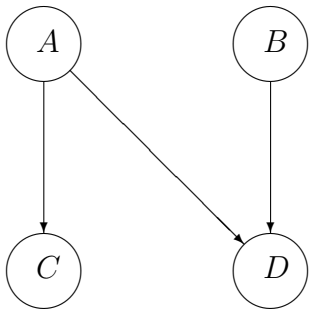


Figure 7.1: Directed Acyclic Graph representing equation 7.1

One of the benefits of using Bayesian modelling here is that prior information, or real-world knowledge, can identify edges which have a potential direct relationship, therefore removing any inappropriate edges and complicating the model unnecessarily.

[1]A cycle occurs when nodes are joined in a closed ring by a set of arrows which are all orientated in the same direction.

7.3 Drawing Bayesian Networks

In this chapter we use a Bayesian Network to represent the combination of models used throughout this thesis. Some standard conventions have been used to depict certain relationships:

- stochastic relationships such as $\beta_{cr} \sim N(\beta_0, \sigma_\beta^2)$ are represented with oval nodes
- logical relationships such as $z_{c+} = \sum_{r=1}^{R} z_{cr}$ are represented as diamond nodes
- constant values such as data and fixed priors are represented as rectangular nodes.

We use the Graphical Modelling structure to show how the variability propagates through the models. It is also a convenient model to show independence and conditional independence relationships.

7.4 Bayesian Network structure

In Chapters 4, 5 and 6 we have generated three models which encompass exposure, accident rate and accident severity data. We incorporate uncertainty at each step and propagate the uncertainty from the previous steps through the models. Figure 7.2 presents the three probabilistic models as a Bayesian Network using the structural rules described in Section 7.3. This diagram not only gives an overview of the whole process but also allows easy identification of dependent and conditionally independent nodes.

The top part of the diagram represents the exposure model described in Section 4.5 with known data z_{cy} (registered number of cars by car type c and year y), e_{cr} (induced exposure by car type c and road type r) and x_{+yr} (number of car km

by road type r and year y) and unknown parameters λ, α_r, β_{cr} and τ used to estimate the disaggregated exposure measure x_{cyr}. We observe x_{+yr}, the sum of traffic over car type c, and use this to restrict the posterior for x_{cyr}.

$$\log \boldsymbol{x} \sim N(\boldsymbol{\beta} + \log \boldsymbol{z}, \tau^2)$$
$$\log \boldsymbol{e} \sim N(\boldsymbol{\alpha} + \log \boldsymbol{z}^+ + \boldsymbol{\beta}, \lambda^2)$$

To the right of this part, the exposure distribution for age of car is introduced, as discussed in Section 4.7, with known data a_{cyg} and some associated uncertainty from δ^2.

$$\log \boldsymbol{a} \sim N(\boldsymbol{\theta}, \delta^2)$$

where $\log a_{cyg}$ is the log of the number of registered vehicles by year (y), car type (c) and car age (g) given an uninformative prior on θ_{cyg}:

$$\theta_{cyg} \sim N(\theta_0, \sigma_\theta^2)$$

where $\boldsymbol{\theta_0} = 0$ and $\sigma_\theta^2 = 10$, and an a priori relative error of 25%:

$$\delta^2 \sim IG(\delta_a, \delta_b)$$

where $\delta_a = 4$ and $\delta_b = 0.05$

We then combine x_{cyr} and θ_{cyg} together to produce the overall traffic exposure measure t_{cygr} used in the accident rate model in Chapter 5. It is clear from the structure of this part of the network that, due to data limitations, in the exposure data, car age and road type are conditionally independent given year and car type. This limitation is discussed in Section 4.7.

$$t_{cygr} = \frac{\theta_{cyg}}{\sum_g \theta_{cyg}} \times x_{cyr}$$

In the accident rate model, accidents are assumed to be Poisson distributed with some overdispersion and are modelled in a generalised linear model with the simulated exposure data t_{cygr} as an offset.

$$S_{cygrb} \sim Po(\chi_{cygrb})$$

$$\log(\chi_{cygrb}) = \log(t_{cygr}) + \gamma_0 + \gamma_1[c] + \gamma_2 y + \gamma_3[g] + \gamma_4[r] + \gamma_5 b + \cdots + \psi_{cygrb}$$

The parameters are given uninformative priors $\gamma_j \sim N(0, 100)$, $\sigma^2_\psi \sim IG(0.01, 0.01)$ and $\boldsymbol{\psi} \sim N(0, \sigma^2_\psi)$ is the overdispersion parameter.

Finally, the bottom row of the diagram represents the accident severity model discussed in Chapter 6 which estimates the proportion p of all single vehicle accidents where a car occupant is killed or seriously injured (S) which are fatal (F).

$$F_i \sim Bin(p_i, S_i)$$
$$\text{logit}(p_i) = \sum_j \phi_j X_{ij}$$

where the unknown coefficients have uninformative priors $\phi_j \sim N(0, 100)$.

7.5 Prediction

7.5.1 Exposure model

We use the structure of the Bayesian Network to predict accident rates and severity proportions in 2011. Initially we estimate the disaggregated exposure data from the first two parts of the model. Figure 7.2 shows that the direct inputs into t_{cygr} are θ_{cyg} and x_{cyr} and their posterior distributions can be derived directly from linked nodes τ^2, z_{cy} and β_{cr} for x_{cyr}, and a_{cyg} and δ^2 for θ_{cyg}. In this case we know z_{cy}, a_{cyg} and x_{+yr} in 2011, however future years could be predicted by making

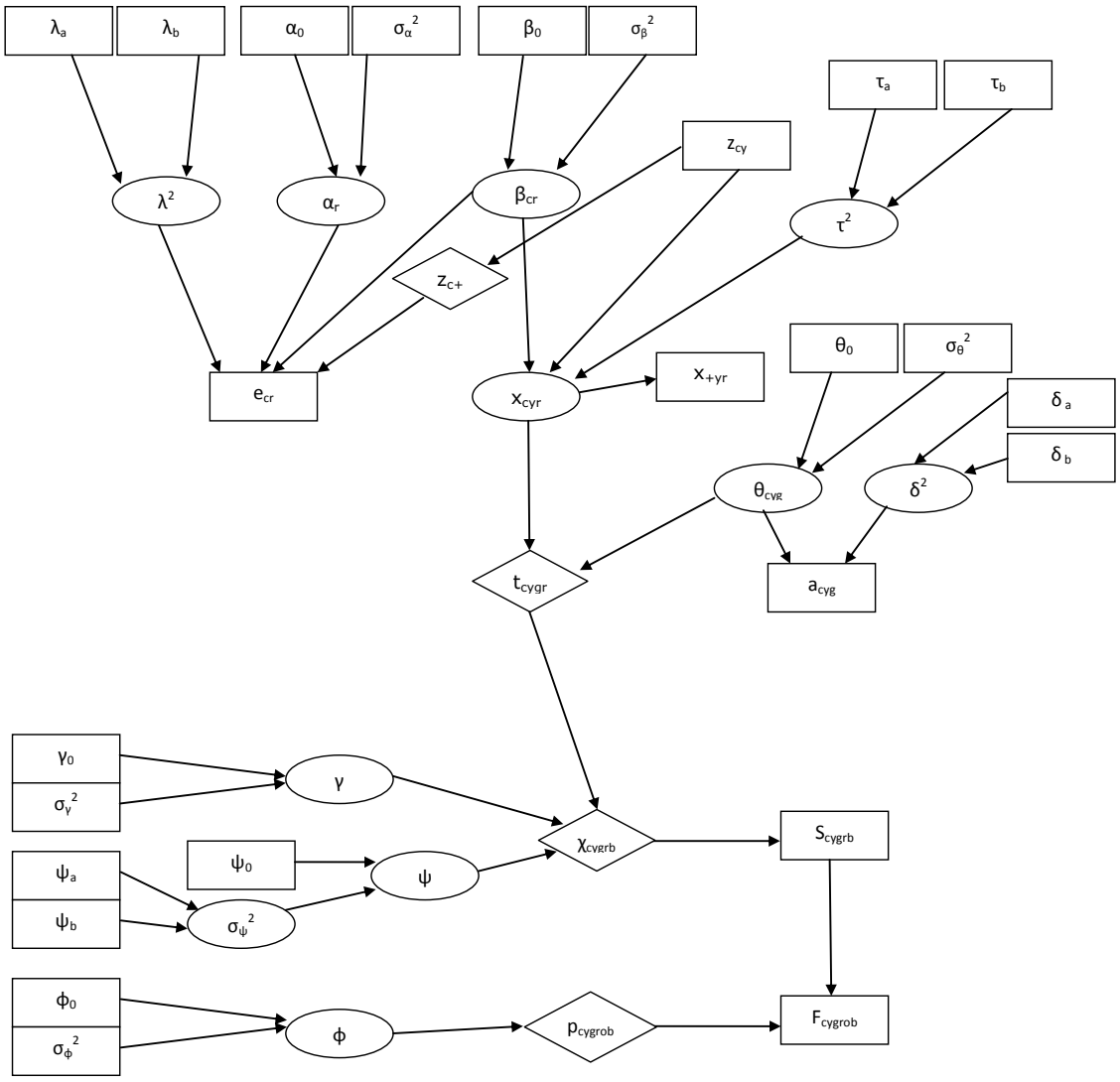


Figure 7.2: Bayesian Network representing exposure, accident rate and accident severity models

assumptions about these datasets. For example we could assume that there was a proportional change equivalent to the proportional change in previous years or no change from the previous year. We have used the known data and extracted a sample of the previously generated posterior distribution for coefficients β_{cr}, τ and δ to generate an exposure distribution for 2011 as shown, compared to 2010 and 2009, in Figure 7.3.

This shows that the uncertainty around the 2011 results is larger than those in 2009 and 2010 as the model is based on these earlier years. There is some suggestion of an increase in small saloons and a decrease in medium saloons in 2011 and overall there has been a slight increase in the amount of traffic in 2011 compared to 2010 (387 billion car km compared to 386 billion car km in 2010), but it has not returned to the peak of 394 billion car km seen in 2009.

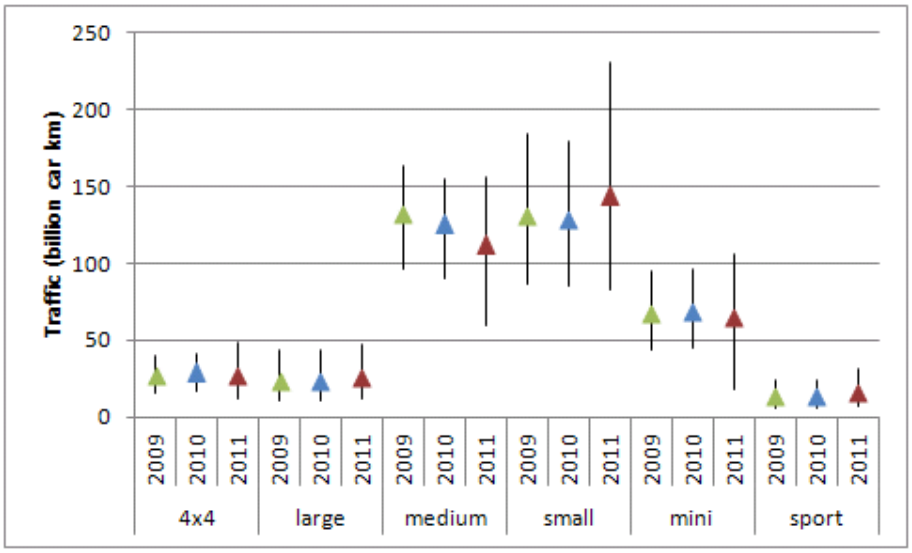


Figure 7.3: Modelled mean exposure and 95% posterior intervals by car type from 2009 – 2011

7.5.2 Accident rate model

The uncertainty in the exposure data is carried forward into the accident rate model. We take draws from the posterior distributions of coefficients γ_j from the tight prior exposure model shown in Section 5.4.3 and add the log of the predicted exposure to get an estimate for accident numbers S in 2011. The model assumes accidents are Poisson distributed with mean χ and overdispersed variance where σ_ϕ^2 is approximately 50, as shown in Tables D.2 and D.3. Figure 7.4 shows the medians[2] and posterior intervals of the simulated counts for variable exposure values using the reduced variability. The model predicts that there was a rise in small and large saloon accidents and accidents involving sports cars, and a reduction in mini, medium saloon and 4x4 accidents. The overall model median predicts a slight decrease in the total number of accidents in 2011 from 3 002 in 2010 to 2 996 in 2011. The actual numbers are not quite as high as predicted: in 2010 there were 2 754 single vehicle car accidents where the car occupant was killed or seriously injured and, in 2011, this decreased to 2 591. This 2011 value falls within the posterior interval for the year: (2 327, 4 771).

The equivalent modelled and predicted counts for 2009 – 2011 with no variability in the exposure data are shown in Figure 7.5. These posterior distributions are smaller and more symmetrical and the medians for each year are closer to the actual counts. The model predicts a small decrease in the number of accidents from 2 908 in 2010 to 2 900 in 2011.

7.5.3 Accident severity model

Using the predicted total number of fatal and serious accidents from Section 7.5.2 and the economy model derived in Chapter 6 we can predict the number of fatal

[2]As the prediction intervals are asymmetric we use the median value as the most appropriate overall summary statistics.

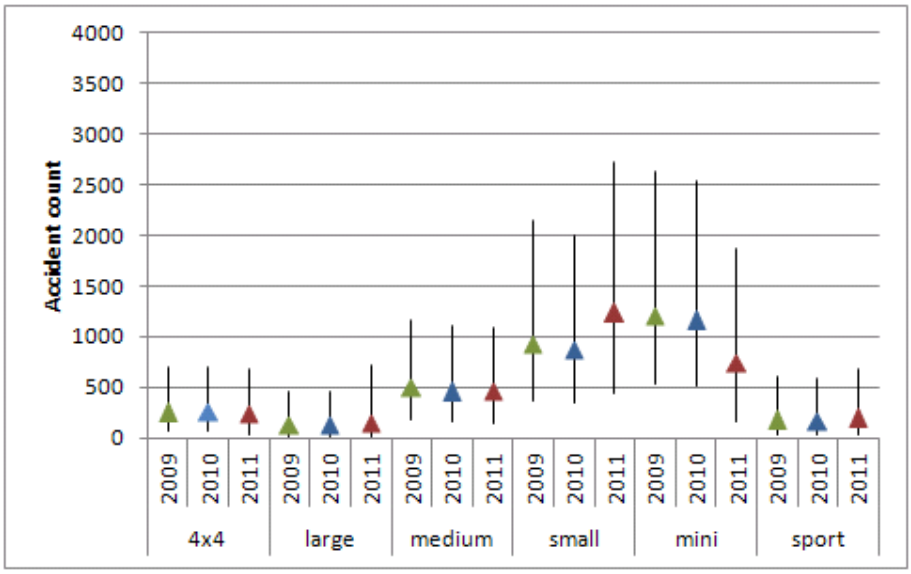


Figure 7.4: Modelled median and 95% posterior intervals around fatal and serious accident counts by car type for 2009 – 2011 with variable exposure from Figure 7.3

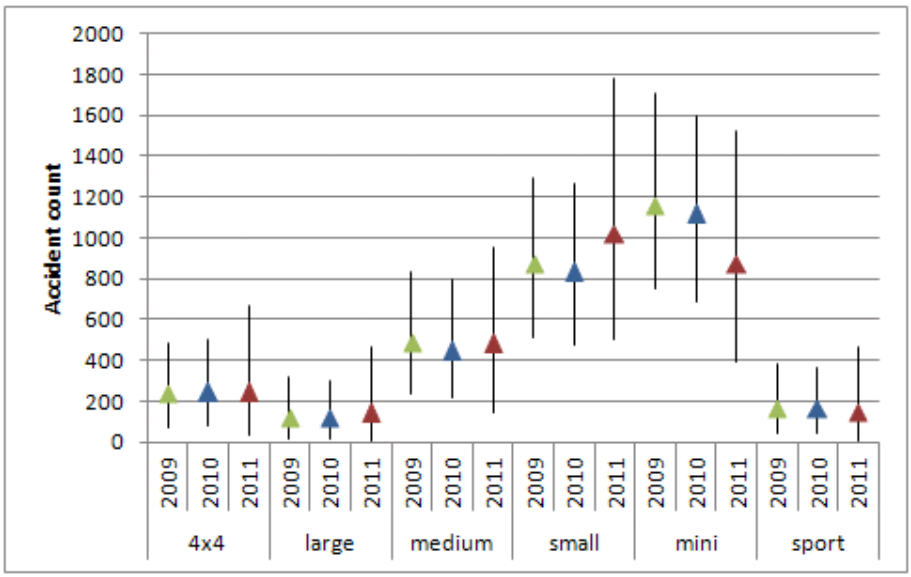


Figure 7.5: Modelled median and 95% posterior intervals around fatal and serious accident counts by car type for 2009 – 2011 with fixed exposure

accidents in 2011. The model assumes that fatal accidents are distributed with a Binomial distribution with n equal to the number of fatal and serious accidents and a severity proportion p derived from a logit function with coefficients ϕ_j derived in Section 6.3. At the time of writing the GDP per capita (the measure of economy used in the model) was not available for 2011. We have therefore made an estimate of what the GDP value may be with two scenarios. Firstly that there was no change from 2010 to 2011 and secondly that the GDP increased at the same rate from 2010 to 2011 as it did from 2009 to 2010. The two scenarios lead to estimated economy values of £23 455 (scenario 1) and £24 388 (scenario 2) per capita respectively.

Figure 7.6 contains the predicted number of fatal accidents in 2011 based on the number of accidents predicted in Section 7.5.2 under the two GDP scenarios, and the modelled number of fatal accidents in 2009 – 2010, based on the modelled number of accidents shown in Figure 7.4. The predicted median total number of fatal accidents in 2011 is slightly higher than that in 2010: 495 fatal accidents in 2010 compared to 517 (scenario 1) and 529 (scenario 2) in 2011. Most of the increase is seen in the small saloon category with an associated decrease in minis, matching the increase and decrease in the traffic and all accidents in small saloons and minis shown in Figures 7.3 and 7.4. The remaining car types predict similar medians in 2011 to those 2010 with marginally larger confidence intervals in general.

In fact there were 307 single car accidents in 2011 where a car occupant was killed. This is considerably fewer than predicted in the model due to the accident rate model predicting significantly more accidents overall. The proportions of the actual accident numbers that were fatal in 2010 and 2011 were approximately 10% and 12% respectively, and the median predicted proportions were both 16%, considerably higher than the actual proportions. Once again the prediction intervals are large and raise questions about their practical use. We discuss this in Section

8.2.

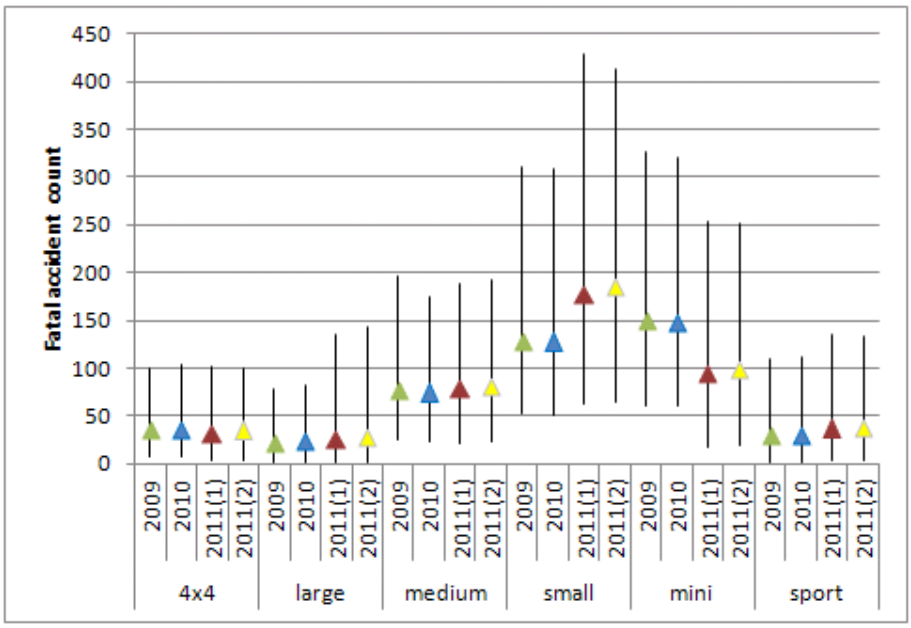


Figure 7.6: Modelled median and 95% posterior intervals for fatal accident counts by car type for 2009 – 2011

Chapter 8

Conclusions

8.1 Summary

In the last few years around 2 000 people each year have died as the result of a road accident in Great Britain. The trend in fatal casualties is closely monitored by the Department for Transport and Government road safety policy is guided by these trends. This pattern in fatal road accidents has fluctuated in recent years after many years of a continual steady downwards trend. The steady decline was replaced with a period of very little reduction from 1994 to 2006 followed by a sharp drop in fatalities in 2007 to 2010. From the mid 1990s several types of fatal accidents influenced the stationary trend – in particular, there was a rise in the number and severity of single car accidents, fatal accidents involving large cars such as 4x4s and fatal accidents involving older cars. We have investigated these accident types by modelling the trends in single vehicle accidents where a car occupant was killed or seriously injured.

In order to accurately reflect changes in the types, sizes and age of vehicles using the road network over time it has been necessary to include exposure data in the

modelling. Exposure data measures the exposure to the risk of being involved in a road accident, the most common of which is the measure of traffic measured in vehicle kilometres. Alternatives to this exposure measure include population, vehicles registered and induced exposure data based on drivers involved but not at fault in an accident. All have their limitations and each of these measures has some associated uncertainty. For example, the traffic data (a combination of automatic and manual surveys) can be disaggregated by road type, year and vehicle type but cannot identify different car types or ages or different driver characteristics; use of the registered vehicle data as an exposure measure allows disaggregation by car type and age but assumes that the distance travelled on the road network is the same for each vehicle; and induced exposure makes the questionable assumptions that drivers not at fault in an accident are a random representation of the vehicles on the road at the time of the accident, and that it is possible to tell which driver was at fault in the accident.

The traffic data is the preferred basis for road accident exposure data as it is the best proxy that is known for exposure to accident risk. There is, however, limited disaggregation and too much variability within each of the disaggregated categories of exposure for it to be satisfactory on its own: different car types are known to be used on different road types at different times of day for example and this is not captured in the traffic data.

There are a number of ways of reducing internal variability, be it variability across time, space or between individuals. These have not been discussed in the road safety literature, but are occasionally demonstrated in epidemiology research. It is possible to include variability as an input in a Bayesian modelling framework; alternatively some additional data can be used – an example of this is to combine a sample survey of detailed information with much broader but less detailed central site measurements. We have reduced the variability in traffic exposure by introducing information from two other sources of exposure (the registered vehi-

cle data and induced exposure data). With these three sources combined we can disaggregate exposure to a greater extent than possible before and incorporate some vital variability from the factors car type, road type, year and car age. We recognise that other forms of variability remain, such as driving habits of young and older drivers, and these are a source of uncertainty.

Two of the exposure data sources are based on sample surveys and therefore also have some associated internal uncertainty. In the past accident risk analysis has not taken into account this uncertainty, assuming that exposure values are fixed and true. The disadvantage of this approach is that it often leads to overconfidence in results and a lack of clear understanding from the results of where improvements in data collection are required. We use a Bayesian analysis for an explicit way to include these multiple sources of uncertainty.

The final exposure model alleviates some of the problems of variability and uncertainty by combining the three sources of exposure using a probabilistic log-Normal model with model priors representing our uncertainty in each data source. This allows further disaggregation of the measure of traffic by car type and car age and strengthens the ability to draw robust conclusions. This is the first time that road accident exposure measures have been combined in order to allow further disaggregation.

This exposure model has shown that:

- car traffic has in general risen over the last 12 years;
- traffic levels are highest on Minor roads, followed by A roads and Motorways;
- small cars (encompassing small and medium saloons, superminis and minis) contribute most of the car traffic, and the gap between these cars and larger cars is bigger on Minor roads than A and Motorways;
- cars aged 6-10 years contribute the most traffic on all road types, and the proportional contribution of this age group has increased in recent years as

the amount of traffic from newer cars has decreased; and

- there has been considerable growth in the amount of 4x4 and people carrier traffic and mini and supermini traffic over the last 12 years on all road types.

The main purpose for exposure data in road accident analysis is to evaluate whether accident numbers for a particular group of drivers are high or low relative to how often a particular scenario occurs – a large number of accidents may suggest a large number of drivers exposed to a certain scenario, or may identify a particularly risky situation. With the simulated exposure measure derived for car type and age it was possible to establish the accident rate, or relative risk, over different car types, car ages, road types and years. The number of fatal accidents is insufficient to produce reliable models so the accident rate models are based on single car accidents where a car occupant was fatally or seriously injured. Accidents were assumed to be Poisson distributed and have been modelled in a generalised linear model with exposure as an offset over five main effects: year, car type, car age, road type and the binary factor which distinguishes accident at or not at a bend, all factors which had been identified as influential in the changing trend in fatal accidents over the last twelve years. A Bayesian model selection process identified two models with several two way interactions including car type and road type, car age and road type and car age by year. The model shows the following patterns:

- minis & superminis, and sports cars have higher accident rates than other car types;
- older cars have higher accident rates than newer cars;
- accidents occurring on Minor roads are more likely relative to the number of vehicle kilometres travelled on Minor roads than other road types;
- accident rates have reduced over time;
- the difference in rates between older cars and newer cars was much greater in 1999 than 2010, most likely due to improvements in secondary safety; and

- accident rates for minis and large saloons on Minor roads are considerably higher than those on A roads but for medium saloons the rates are equal.

In order to assess the trend in fatal accidents, a severity model was derived to estimate the fatality proportion of all accidents where a car occupant was killed or seriously injured. A Binomial model uses the logit link to regress on each of the main effects used in the accident rate model plus an additional main effect reporting whether the car overturned. The model assumes all the main effects are required and a model selection procedure selects three models with the interactions age and road, and overturn and bend and the two together to adequately model the severity rates. The results of these models have been averaged and show that:

- minis and superminis, 4x4 and people carriers, and small saloons have lower severity rates than other car types;
- accidents occurring on Minor roads are generally less severe than other road types;
- accidents where the car did not overturn led to less severe injuries; and
- accidents involving an old car (over 15 years old) were more likely to be fatal accidents.

Accident severity is known to be reduced in younger and bigger cars due to secondary safety developments. Small cars may be resulting in less severe accidents due the types of drivers using these cars.

The combination of likelihood and severity is important – highly likely and severe accidents are high priority for governments to influence. In general the models show that highly likely accidents have a lower severity proportion, for example minis have a considerably higher accident rate but a low severity proportion relative to other car types. Old cars are the exception – they are relatively more likely to be involved in a serious or fatal accident, and result in injuries which are more severe.

We introduced a measure of economy into the accident rate and accident severity models in place of the factor year and showed that, for the severity model only, economy is a better predictor than year. It is difficult to show definitively but this suggests that the recession may have affected the number of fatal accidents but had less or no influence on the total number of serious and fatal accidents. It is most likely that it is an indirect effect; we have observed that there has been a reduction in young drivers and average speeds over the period where the economy was declining which are both likely to have an effect on severity.

The three models of exposure, accident rates and accident severity have been combined to predict accident rates and severity proportions in 2011. Initially a disaggregation of the exposure data was estimated from known data and a sample of the previously generated posterior distribution for coefficients of the exposure model. This showed a slight increase in traffic since 2011, with an increase in small saloon traffic and a corresponding decrease in medium saloon traffic. These predicted exposure values were combined with the previously generated posterior distribution for coefficients in the accident rate model and predicted that there was an increase in small and large saloons single car accidents and a decline in the number of single car accidents involving minis and medium saloons in 2011. The model medians appear to over-predict the total number of accidents in 2011, but the actual figure was within the 95% posterior intervals predicted. This over-prediction is carried forward into the accident severity model where, based on the previously generated posterior distributions for coefficients in the severity model and predictions from the accident rate model, we predict a severity rate between 16% and 17%, considerably higher than the actual 12% of fatal and serious accidents which were fatal. These differences suggest that further information is required in these models, as discussed in Section 8.2.

In summary, we have shown that it is possible to include uncertainty in measures of exposure for road accidents, and propagate that uncertainty through to accident

rate and severity models. It is also possible, and indeed informative, although it has not been done before, to combine certain exposure datasets together to enable more disaggregated accident rates to be modelled. We have established the first statistical suggestion that the economy influences road accident severity, with a recession resulting in a lower number of fatal accidents.

8.2 Limitations and further work

Throughout this thesis the derived models have been based on single vehicle car accidents in which an occupant was killed or seriously injured. Whilst this is a particular accident type of concern it only encompasses around 12% of all accidents involving a fatal or serious injury and around 1% of all reported accidents. Results cannot be generalised directly for other vehicle types or accident types and this is a limitation in the results. However, if we believed, and most research in this area does so, that the measure of traffic was a decent proxy for exposure to multi-vehicle accidents then it is a relatively simple task to apply the simulated exposure data from Chapter 4 to accidents involving more than one car and derive models for accident rate and severity in multi-car accidents. As discussed in Section 2.2.2, in theory, the concept of exposure for multi-vehicle accidents is slightly different: there must be, by definition, more than one vehicle present for there to be a multi-vehicle accident, and therefore the ideal measure of exposure must be time when or distance over which there is more than one vehicle present at any point on the road. In general when multi-vehicle accident rate modelling is reported, it is based on large groups of accidents and it may be a reasonable assumption to say that measures of traffic are the best proxy to 'multiple vehicles present'. Once accidents are disaggregated into more factors, as we have here, the link between true exposure for multi-vehicle accidents and traffic is less obvious and requires more research in the future.

The technique used in Chapter 4 to derive the exposure data for cars could be applied to other vehicle exposure data and the accident rate modelling could then be extended to various vehicle types. Accidents involving pedestrians are more difficult and it requires further work to derive a useful measure of exposure for pedestrians in road accidents. Current practice is to use either vehicle traffic or estimates of pedestrian traffic and neither assess time when a pedestrian and a vehicle are present at a particular point.

In summary, there are limitations in using traffic as an exposure measure for accident rate modelling, particularly for multi-vehicle and pedestrian accidents, however it is the best proxy currently available and disaggregating these data using further sources of exposure data makes results and conclusions more robust by reducing the variability within the estimates.

There are additional sources of exposure data which would allow the exposure measure to be disaggregated further, such as the age and gender of drivers and the number of occupants in different car types. In this context the age of drivers is of particular interest as one hypothesis questions whether the recent recession has reduced the number of young drivers being able to afford to take their driving test or being able to afford to drive once they have passed their test, thus reducing the number of young drivers (known to be more accident prone in general) driving on the road network. These data are available from sources such as the National Travel Survey and could be included in the exposure measure using similar methods to those described above. They have not been included in this research as we concentrate on vehicle rather than occupant exposure. In practice these additional factors would increase the size of the model so much that computational difficulties would most likely become the overriding issue and any results could be even less certain that currently. It would be interesting to disaggregate the original traffic measure x_{yr} into driver characteristics rather than vehicle characteristics using the aforementioned dataset to see if any particular

factors were strong predictors in accident rate and severity models.

Predicting forward was a theme in Chapter 7, using coefficients from the models based on 1999 – 2010 data and propagating uncertainty through each step. The predictions of number of fatal and serious accidents were highly uncertain, in one case for minis a range from 250 to 3500 accidents was predicted for 2011. The maximum of this range is 30% bigger than the total number of single car accidents in 2010 for all car types combined. Generally, year by year these numbers change by a maximum of +/-30% and therefore the ranges are unrealistic. This implies that specified priors have perhaps been too weak and further work is required to hone these down to realistic values. It also brings up the question of the use of statistical uncertainty in prediction intervals for road safety policy purposes. If the object of interest in a prediction model is how much the accident rate is going to change from year to year, for example, then propagating uncertainty in the exposure measure, which is likely to be similar from year to year, will increase the prediction range unnecessarily. The aim is often not to get an assessment of the possible range of accident rates taking into account that the exposure measure is not 100% accurate but to accept limitations in the exposure data and assess the likely change in reported accident rates in future years without this added annual uncertainty. The statistical prediction intervals that have been derived in Section 7.5 are not useful for the purposes of allocating Government budgets, defining policies for road safety and prioritising interventions for specific user groups – some work is needed to assess how much of the variability that has been carried through the modelling is required for practical use.

Appendix A

Induced exposure

Assumption 1: not at fault drivers are a random sample of the local driving population

The major assumption made in using induced exposure methods is that not at fault drivers are a random sample from the driving population. Several reports assess this assumption in different ways.

Assessing assumption 1: Stringent criterion for AF/NAF

It is suggested by af Wahlburg and Dorn (2007) that this assumption is heavily dependent on how stringent the criterion for responsibility is. Within the not at fault driver sample there are likely to be some drivers who were partly at fault in their accident, but not identified by the investigation. This group could include not at fault young drivers, due to their lack of defensive driving skills and thus they could also be considered partly AF. Including these drivers in the not at fault group affects the estimation of exposure and violates the assumption that the not at fault groups are a random sample of the driving population.

Analysis of bus driver statistics from the UK and Sweden suggests that using a

stringent criterion for culpability (i.e. in Sweden) produces not at fault exposure data comparable to non-accident drivers, and in the UK (with a less stringent culpability criterion) the not at fault drivers are not such a random sample of the general bus driving population.

It is suggested that a larger study is required across further driver types to be able to generalise to the whole driving population in part due to the fact that culpability varies by vehicle type. These levels of culpability should then be used to determine how stringent these criteria should be.

Chandraratna and Stamatiadis (2009) complete a similar evaluation to af Wahlburg and Dorn (2007) of the not at fault assumption which takes two subsets of drivers involved in crashes detailed in the Kentucky accident database from 1995-1999. The first subset is the second driver (defined as NAF) in a multi-vehicle accident, the second subset is any drivers after the first and second drivers.

It is hypothesised that the assumption of 'drawn from the exposed driving population' is more appropriate for the second sample than the first.

The first sample will probably include some partly at fault drivers which biases the results. A few patterns (in particular left turn crashes and rear end crashes where investigators may find it difficult to assess at fault and NAF) show significant differences between sample 1 and sample 2, suggesting two different driving populations, but this could be due to misidentification of NAF.

Gender and age were also tested and showed no significant differences between the two samples suggesting that, for most crash types, indeed for accidents where it is 'easy' to determine AF/NAF, the quasi-induced exposure method is valid.

Lardelli-Claret et al. (2005) compare two quasi-induced exposure methods to investigate biases produced by the classical quasi-induced exposure method. Firstly classical quasi-induced exposure method compared at fault and not at fault drivers

in approximately 450 000 multi-vehicle accident and single vehicle accident in Spain from 1993-2002. Using Multinomial logistic regression, two odds ratios: involvement ratio (IR) not at fault v. IR at fault multi-vehicle accident and IR not at fault v. IR at fault single vehicle accident were calculated for each category of drivers defined by several factors. Difference between the results were detected using a χ^2 test.

Secondly, only clean (well defined uniquely responsible driver for each accident) two-vehicle collisions were analysed in a paired-by-collision analysis. Conditional logistic regression compared characteristics between groups and an odds ratio (OR) for each driver category was estimated.

For the two methods, the OR were similar for almost all driver and vehicle related factors. However, the ORs in the paired sample analysis were more than 10% higher for psychophysical conditions.

Theoretically of the two methods, the paired method allows for better control of measured and unmeasured environmental factors, however the results do not differ sufficiently to reject the classical quasi-induced exposure method and both methods are suggested appropriate for estimating IR of two-vehicle collisions.

Assessing assumption 1: Biased representation of risky circumstances

Biases can exist for drivers travelling in more risky locations or at more risky times. In Stamatiadis and Deacon (1996), disaggregated data appears to provide a much better estimate of the AR driving population than highly aggregated data. This also gives an idea of confounding factors in the exposure data.

Assessing assumption 1: Speed bias

Jiang and Lyles (2007) investigate another type of bias in the quasi-induced exposure method. They propose that for vehicles that routinely travel faster, the

involvement ratio will be underestimated and vice versa. The empirical example shows that for faster vehicles, the IR increases with speed limit and for lower vehicles, IR decreases with speed limit. It is however, not possible to tell whether this is affected by the increasing speed differential or the change in driving population on these different roads. The overall conclusion is that where a high population of slow moving vehicles exist, the speed differential may bias the results from the quasi-induced exposure method.

Bias can be assessed by examining a table of NAF/AF by speed differential and no speed differential, and comparing marginal totals (Jiang and Lyles 2007, see table 1 in).

Assumption 2: the characteristics of not at fault drivers in multi-vehicle accidents is the same as that in single vehicle accidents

Accident propensity (the ratio of the proportion of one group in the at fault drivers distribution compared to not at fault drivers) can also be calculated for single vehicle accidents replacing the at fault proportion with the proportion in single vehicle accidents, if you are willing to assume that the extent of exposure to accidents is the same for single vehicle accidents and multi-vehicle accidents, i.e. the distribution of drivers/vehicles at fault in multi-vehicle accidents is the same as in single vehicle accidents.

Stamatiadis and Deacon (1996) estimated exposure for driver age group and vehicle types separately, for single vehicle accidents and multi-vehicle accidents. Comparing their data (with conventional exposure data) suggests that the distribution of driving groups and vehicle types is not the same for single vehicle accidents and multi-vehicle accidents so data should not be combined.

Conclusions

A series of conclusions, some contradictory, can be drawn from the papers re-

viewed:

- quasi-induced exposure improves on induced exposure techniques as responsible drivers in single vehicle accidents and multi-vehicle accidents are not assumed to be from the same population (Lyles et al. 1991; Stamatiadis and Deacon 1996);
- bias in defining driver responsibility should not affect results dramatically (Lyles et al. 1991);
- for some crash types, where is it more difficult to determine which driver is responsible, the not at fault group will not be representative of the exposed driving population (Chandraratna and Stamatiadis 2009);
- inexperienced drivers are over represented in the not at fault group (Stamatiadis and Deacon 1996);
- not at fault drivers seem to be approximately a random sample from the general driving population (Lyles et al. 1991);
- validation of this technique is no more challenging that other techniques and quasi-induced exposure data are an important improvement over other less readily available data (Lyles et al. 1991);
- disaggregated data will help to avoid bias (Stamatiadis and Deacon 1996);
- in cases where there are a high proportion of slow moving vehicles, a biased not at fault group may exist (Jiang and Lyles 2007).

Lyles (1994) compiled a list of when the quasi-induced exposure method should not be used:

- where the relative risks are insufficient and specific accident rates calculations are required;
- where the not at fault distribution is known to be biased;
- where a small sample size exists;
- when data are not cleaned (hit and run removed etc.).

Appendix B

Posterior distributions for exposure modelling

B.1 Early models

B.1.1 Model 1

$$
\begin{aligned}
p(x \mid z) &\propto p(z \mid x)p(x) \\
&\propto \exp\Big(-\frac{1}{2\tau^2}(z-\beta Ax)^T(z-\beta Ax) - \frac{1}{2\sigma^2}(x-\mu)^T(x-\mu)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big\{\tau^{-2}\big[z^Tz - \beta(Ax)^Tz - z^T\beta Ax + \beta^2(Ax)^TAx\big] \\
&\qquad + \sigma^{-2}\big[x^Tx - \mu^Tx - x^T\mu + \mu^T\mu\big]\Big\}\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big\{x^T\Big[\frac{\beta^2}{\tau^2}A^TA + I\sigma^{-2}\Big]x + x^T\Big[-\frac{\beta A^Tz}{\tau^2} - \frac{\mu}{\sigma^2}\Big] \\
&\qquad + \Big[-\frac{\beta z^TA}{\tau^2} - \frac{\mu^T}{\sigma^2}\Big]x\Big\}\Big)
\end{aligned}
$$

$$\exp\big(-\frac{1}{2}(x-\theta)^T\Sigma^{-1}(x-\theta)\big) \equiv \exp\big(-\frac{1}{2}x^T\Sigma^{-1}x - \theta^T\Sigma^{-1}x - x^T\Sigma^{-1}\theta + \theta^T\Sigma^{-1}\theta\big)$$

$$\begin{aligned}
\Sigma^{-1} &= \frac{\beta^2}{\tau^2}A^TA + I\sigma^{-2} \\
\Sigma^{-1}\theta &= \frac{\beta A^Tz}{\tau^2} + \frac{\mu}{\sigma^2} \\
\theta &= \Big(\frac{\beta^2}{\tau^2}A^TA + I\sigma^{-2}\Big)^{-1}\Big(\frac{\beta A^Tz}{\tau^2} + \frac{\mu}{\sigma^2}\Big)
\end{aligned}$$

$x \mid z \sim N(\theta, \Sigma)$

where

θ is a $C \times Y \times R$ vector

Σ is a $C \times Y \times R$ by $C \times Y \times R$ covariance matrix

Conditional posteriors for parameters (non deterministic)

Beta

$\beta \sim N(\beta_0, \sigma_\beta^2)$

$$\begin{aligned}
p(\beta \mid z) &\propto \exp\Big\{-\frac{\tau^{-2}}{2}(z-\beta Ax)^T(z-\beta Ax)\Big\}\exp\Big\{-\frac{\sigma_\beta^{-2}}{2}(\beta-\beta_0)^2\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\big[\tau^{-2}(z^Tz + \beta^2(Ax)^TAx - 2\beta z^TAx) + \sigma_\beta^{-2}(\beta^2 - 2\beta_0\beta + \beta_0^2)\big]\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\big[\beta^2(x^TA^TAx\tau^{-2} + \sigma_\beta^{-2}) - 2\beta(\tau^{-2}z^TAx + \sigma_\beta^{-2}\beta_0) + ...\big]\Big\}
\end{aligned}$$

$\beta \mid z \sim N(\frac{\tau^{-2}z^TAx + \sigma_\beta^{-2}\beta_0}{\tau^{-2}x^TA^TAx + \sigma_\beta^{-2}}, \tau^{-2}x^TA^TAx + \sigma_\beta^{-2})$

Mu

$\mu \sim N(\mu_0, \sigma_0^2)$

$$
\begin{aligned}
p(\mu \mid x) &\propto (\sigma^2)^{-\frac{n}{2}} (\sigma_0^2)^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\Big[\sigma^{-2}(x-\mu)^T(x-\mu) + \sigma_0^{-2}(\mu-\mu_0)^T(\mu-\mu_0)\Big]\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\sigma^{-2}(x^Tx - 2\mu^Tx + \mu^T\mu) + \sigma_0^{-2}(\mu^T\mu - 2\mu_0\mu + \mu_0^2)\Big]\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[(\sigma_0^{-2} + \sigma^{-2})\Big(\mu - \frac{x+\mu_0}{\sigma_0^{-2}+\sigma^{-2}}\Big)^T\Big(\mu - \frac{x+\mu_0}{\sigma_0^{-2}+\sigma^{-2}}\Big) + ...\Big]\Big\}
\end{aligned}
$$

$\mu \mid x \sim N(\frac{\sigma^{-2}x + \sigma_0^{-2}\mu_0}{\sigma_0^{-2}+\sigma^{-2}}, \sigma_0^{-2} + \sigma^{-2})$

Tau

$\tau^{-2} \sim \Gamma(\alpha_\tau, \beta_\tau)$

$$
\begin{aligned}
p(\tau^{-2} \mid z) &\propto p(z \mid \tau^{-2}) p(\tau^{-2}) \\
&\propto (2\pi)^{-\frac{m}{2}} (\tau^{-2})^{\frac{m}{2}} \exp\Big(-\frac{\tau^{-2}}{2}(z-\beta Ax)^T(z-\beta Ax) - \beta_\tau \tau^{-2}\Big) \tau^{-2(\alpha_\tau - 1)} \\
&\propto (\tau^{-2})^{\frac{m}{2}+\alpha_\tau - 1} \exp\Big(-\tau^{-2}\Big[\frac{1}{2}(z-\beta Ax)^T(z-\beta Ax) + \beta_\tau\Big]\Big)
\end{aligned}
$$

$\tau^{-2} \mid z \sim \Gamma(\frac{m}{2} + \alpha_\tau, \frac{1}{2}(z-\beta Ax)^T(z-\beta Ax) + \beta_\tau)$

Sigma

$\sigma^{-2} \sim \Gamma(\alpha_\sigma, \beta_\sigma)$

$$
\begin{aligned}
p(\sigma^{-2} \mid x) &\propto (x \mid \sigma^{-2}) p(\sigma^{-2}) \\
&\propto (\sigma^{-2})^{\frac{n}{2}} \exp\Big(-\frac{\sigma^{-2}}{2}(x-\mu)^T(x-\mu) - \beta_\sigma \sigma^{-2} \Big) \sigma^{-2(\alpha_\sigma - 1)} \\
&\propto (\sigma^{-2})^{\frac{n}{2}+\alpha_\sigma - 1} \exp\Big(-\sigma^{-2} \Big[\frac{1}{2}(x-\mu)^T(x-\mu) \Big] + \beta_\sigma \Big)
\end{aligned}
$$

$\sigma^{-2} \mid x \sim \Gamma(\frac{n}{2} + \alpha_\sigma, \frac{1}{2}(x-\mu)^T(x-\mu) + \beta_\sigma)$

B.2 Introducing induced exposure

B.2.1 Model 2

$$
\begin{aligned}
p(x \mid a = 0, b = \infty) &= \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{x-\mu}{2\sigma^2} \big)}{1 - \Phi(-\frac{\beta_{cr} z_{cy}}{\tau})}
\end{aligned}
$$

B.3 Detailed posterior working

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau^2, \lambda^2 \mid \boldsymbol{z}, \boldsymbol{e}) &\propto \prod_{cyr} p(x_{cyr} \mid \beta_{cr}, \tau^2) \prod_{cr} p(e_{cr} \mid \alpha_r, \beta_{cr}, \lambda^2) \prod_{cr} p(\beta_{cr}) \\
&\quad \prod_{r} p(\alpha_r) p(\tau^2) p(\lambda^2) \\
&\propto \frac{1}{2\pi\tau^2}^{\frac{CYR}{2}} \exp\Big\{ -\frac{1}{2\tau^2} \sum_{cyr} (x_{cyr} - \beta_{cr} z_{cy})^2 \Big\} \\
&\quad \frac{1}{\prod_{cyr} 1 - \Phi(-\frac{\beta_{cr} z_{cy}}{\tau})} \\
&\quad \frac{1}{2\pi\lambda^2}^{\frac{CR}{2}} \exp\Big\{ -\frac{1}{2\lambda^2} \sum_{cr} (e_{cr} - \alpha_r z_{c+} \beta_{cr})^2 \Big\} \\
&\quad \frac{1}{2\pi\sigma_\beta^2}^{\frac{CR}{2}} \exp\Big\{ -\frac{1}{2\sigma_\beta^2} \sum_{cr} (\beta_{cr} - \beta_0)^2 \Big\} \\
&\quad \frac{1}{2\pi\sigma_\alpha^2}^{\frac{R}{2}} \exp\Big\{ -\frac{1}{2\sigma_\alpha^2} \sum_{r} (\alpha_r - \alpha_0)^2 \Big\} \\
&\quad \frac{1}{\tau^2}^{\tau_a + 1} \exp\{-\frac{\tau_b}{\tau^2}\} \frac{1}{\lambda^2}^{\lambda_a + 1} \exp\{-\frac{\lambda_b}{\lambda^2}\}
\end{aligned}
$$

β **posterior**

$$p(\boldsymbol{\beta} \mid \boldsymbol{x}, \tau^2, \lambda^2, \boldsymbol{\alpha}, \boldsymbol{z}, \boldsymbol{e}) \propto \exp\Big\{ -\frac{1}{2\tau^2}\sum_{cyr}(x_{cyr} - \beta_{cr} z_{cy})^2 - \frac{1}{2\lambda^2}\sum_{cr}(e_{cr} - \alpha_r z_{c+}\beta_{cr})^2 - \frac{1}{2\sigma_\beta^2}\sum_{cr}(\beta_{cr} - \beta_0)^2\Big\} \prod_{cyr}\frac{1}{1 - \Phi(-\frac{\beta_{cr} z_{cy}}{\tau})}$$

α **posterior**

$$p(\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \lambda^2, \boldsymbol{z}, \boldsymbol{e}) \propto \exp\Big\{ -\frac{1}{2\lambda^2}\sum_{cr}(e_{cr} - \alpha_r z_{c+}\beta_{cr})^2 - \frac{1}{2\sigma_\alpha^2}\sum_{r}(\alpha_r - \alpha_0)^2\Big\}$$

$$\propto \exp\Big\{ -\frac{1}{2}\Big[\sum_r \alpha_r^2\Big(\frac{\sum_c((z_{c+}\beta_{cr})^2)}{\lambda^2} + \frac{1}{\sigma_\alpha^2}\Big) - 2\sum_r \alpha_r\Big(\frac{\sum_c e_{cr} z_{c+}\beta_{cr}}{\lambda^2} + \frac{\alpha_0}{\sigma_\alpha^2}\Big)\Big]\Big\}$$

$$\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{ze} \sim N(\boldsymbol{\phi}, \boldsymbol{\Omega})$$

where

$$\Omega_{r\times r}^{-1}\phi_r = \frac{\sum_c e_{cr} z_{c+}\beta_{cr}}{\lambda^2} + \frac{\alpha_0}{\sigma_\alpha^2}$$

$$\Omega_{r\times r}^{-1} = \frac{\sum_c((z_c\beta_{cr})^2)}{\lambda^2} + \frac{1}{\sigma_\alpha^2}$$

$$\phi_r = \Omega_{r\times r} \times \Omega_{r\times r}^{-1}\phi_r$$

τ^2 posterior

$$p(\tau^2 \mid \boldsymbol{x}, \boldsymbol{\beta}, \tau_a, \tau_b, \boldsymbol{z}) \propto \frac{1}{2\pi\tau^2}^{\frac{CYR}{2}} \frac{1}{\tau^2}^{\tau_a+1} \exp\Big\{ -\frac{1}{2\tau^2}\sum_{cyr}(x_{cyr} - \beta_{cr}z_{cy})^2 - \frac{\tau_b}{\tau^2}\Big\} \prod_{cyr}\frac{1}{1-\Phi(-\frac{\beta_{cr}z_{cy}}{\tau})}$$

λ^2 posterior

$$\begin{aligned} p(\lambda^2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_a, \lambda_b, \boldsymbol{z}, \boldsymbol{e}) &\propto \frac{1}{2\pi\lambda^2}^{\frac{CR}{2}} \frac{1}{\lambda^2}^{\lambda_a+1} \exp\Big\{ -\frac{1}{2\lambda^2}\sum_{cr}(e_{cr} - \alpha_r z_{c+}\beta_{cr})^2 - \frac{\lambda_b}{\lambda^2}\Big\} \\ &\propto \frac{1}{\lambda^2}^{\left(\frac{CR}{2}+\lambda_a+1\right)} \exp\Big\{ -\frac{1}{\lambda^2}\Big[\frac{1}{2}\sum_{cr}(e_{cr} - \alpha_r z_{c+}\beta_{cr})^2 + \lambda_b\Big]\Big\} \end{aligned}$$

$$\lambda^2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_a, \lambda_b, \boldsymbol{z}, \boldsymbol{e} \sim IG\Big(\frac{CR}{2} + \lambda_a, \frac{1}{2}\sum_{cr}(e_{cr} - \alpha_r z_{c+}\beta_{cr})^2 + \lambda_b\Big)$$

B.3.1 Model 3

$$p(\boldsymbol{x}, \boldsymbol{g}, \boldsymbol{\beta}, \tau^2, \epsilon^2 \mid \boldsymbol{z}, \boldsymbol{e}) = p(\boldsymbol{x} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{z}, \boldsymbol{e}) p(\boldsymbol{g} \mid \boldsymbol{\beta}, \epsilon^2, \boldsymbol{e}) p(\boldsymbol{\beta}) p(\tau^2) p(\epsilon^2)$$

$$\begin{aligned} p(\boldsymbol{\beta} \mid \tau^2, \epsilon^2, \boldsymbol{z}, \boldsymbol{e}) = & \{2\pi\tau^2\}^{-\frac{CYR}{2}} \exp\Big\{ -\frac{1}{2\tau^2} \sum_{cyr} \Big(x_{cyr} - \beta_{cr} z_{cy} \Big)^2 \Big\} \\ & \{(2\pi\epsilon^2)^R |K|\}^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\epsilon^2} \sum_{r} \sum_{i,j=1}^{C-1} {K_{ij}}^{-1} (g_{ri} - \mu_i)(g_{rj} - \mu_j) \Big\} \\ & \{2\pi\sigma_\beta^2\}^{-\frac{CR}{2}} \exp\Big\{ -\frac{1}{2\sigma_\beta^2} \sum_{cr} \Big(\beta_{cr} - \beta_0 \Big)^2 \Big\} \end{aligned}$$

$$\epsilon^2 \mid \boldsymbol{\beta}, \epsilon_a, \epsilon_b, \boldsymbol{z}, \boldsymbol{e} \sim IG\Big(\frac{R}{2} + \epsilon_a, \frac{1}{2} \sum_{r} \sum_{i,j=1}^{C-1} (g_{ri} - \mu_i)[K^{-1}]_{ij} (g_{rj} - \mu_j) + \epsilon_b \Big)$$

B.4 Log-normal model

B.4.1 Model 4

Joint posterior:

$$
\begin{aligned}
&p(\log \boldsymbol{x}, \log \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau^2, \lambda^2 \mid \boldsymbol{z}, \boldsymbol{e}) \\
&\qquad \propto \prod_{cyr} p(\log \boldsymbol{x} \mid \boldsymbol{\beta}, \tau^2, \boldsymbol{z}) p(\log \boldsymbol{e} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda^2, \boldsymbol{z}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\tau^2) p(\lambda^2) \\
&\qquad \propto \Big(\frac{1}{2\pi\tau^2}\Big)^{\frac{CYR}{2}} \prod_{cyr} \exp\Big[-\frac{1}{2\tau^2}\Big(\log x_{cyr} - \{\log z_{cy} + \beta_{cr}\}\Big)^2\Big] \\
&\qquad\quad \cdot \Big(\frac{1}{2\pi\lambda^2}\Big)^{\frac{CR}{2}} \prod_{cr} \exp\Big[-\frac{1}{2\lambda^2}\Big(\log e_{cr} - \{\log z_{c+} + \beta_{cr} + \alpha_r\}\Big)^2\Big] \\
&\qquad\quad \cdot \Big(\frac{1}{2\pi\sigma_\beta^2}\Big)^{\frac{CR}{2}} \prod_{cr} \exp\Big[-\frac{1}{2\sigma_\beta^2}\Big(\beta_{cr} - \beta_0\Big)^2\Big] \\
&\qquad\quad \cdot \Big(\frac{1}{2\pi\sigma_\alpha^2}\Big)^{\frac{R}{2}} \prod_{r} \exp\Big[-\frac{1}{2\sigma_\alpha^2}\Big(\alpha_r - \alpha_0\Big)^2\Big] \\
&\qquad\quad \cdot \Big(\frac{1}{\tau^2}\Big)^{\tau_a+1} \exp\Big[-\frac{\tau_b}{\tau^2}\Big] \Big(\frac{1}{\lambda^2}\Big)^{\lambda_a+1} \exp\Big[-\frac{\lambda_b}{\lambda^2}\Big]
\end{aligned}
$$

β posterior

$$p(\boldsymbol{\beta} \mid \boldsymbol{x}, \tau^2, \lambda^2, \boldsymbol{\alpha}, \boldsymbol{z}, \boldsymbol{e}) \propto \prod_{cyr} \exp\Big[-\frac{1}{2\tau^2}\Big(\log x_{cyr} - \{\log z_{cy} + \beta_{cr}\}\Big)^2\Big]$$
$$. \prod_{cr} \exp\Big[-\frac{1}{2\lambda^2}\Big(\log e_{cr} - \{\log z_{c+} + \beta_{cr} + \alpha_r\}\Big)^2\Big]$$
$$. \prod_{cr} \exp\Big[-\frac{1}{2\sigma_\beta^2}\Big(\beta_{cr} - \beta_0\Big)^2\Big]$$

$$\propto \prod_{cyr} \exp\Big[-\frac{1}{2\tau^2}\Big((\log x_{cyr})^2 + (\log z_{cy})^2 + \beta_{cr}^2 + 2\beta_{cr} \log z_{cy}$$
$$- 2\log x_{cyr} \log z_{cy} - 2\beta_{cr} \log x_{cyr}\Big)\Big]$$
$$. \prod_{cr} \exp\Big[-\frac{1}{2\lambda^2}\Big((\log e_{cr})^2 + (\log z_{c+})^2 + \beta_{cr}^2 + \alpha_r^2$$
$$+ 2\beta_{cr} \log z_{c+} + 2\alpha_r \log z_{c+} + 2\alpha_r\beta_{cr} - 2\log e_{cr} \log z_{c+}$$
$$- 2\beta_{cr} \log e_{cr} - 2\alpha_r \log e_{cr}\Big)\Big]$$
$$. \prod_{cr} \exp\Big[-\frac{1}{2\sigma_\beta^2}\Big(\beta_{cr}^2 - 2\beta_{cr}\beta_0 + \beta_0^2\Big)\Big]$$

$$\propto \exp\Big[-\frac{1}{2} \sum_{cr} \Big(\beta_{cr}^2 \Big\{ \frac{y}{\tau^2} + \frac{1}{\lambda^2} + \frac{1}{\sigma_\beta^2} \Big\}$$
$$- 2\beta \Big\{ \frac{1}{\tau^2} \sum_{y} \Big(\log x_{cyr} - \log z_{cy} \Big) + \frac{\log e_{cr}}{\lambda^2}$$
$$- \frac{\log z_{c+}}{\lambda^2} - \frac{\alpha_r}{\lambda^2} + \frac{\beta_0}{\sigma_\beta^2} \Big\}\Big)\Big]$$

$$\boldsymbol{\beta} \mid \boldsymbol{x}, \tau^2, \lambda^2, \boldsymbol{\alpha}, \boldsymbol{z}, \boldsymbol{e} \sim N(\boldsymbol{\theta}, \boldsymbol{\Delta})$$

where

$$\Delta_{cr}^{-1}\theta_{cr} = \frac{1}{\tau^2}\sum_{y}\Big(\log x_{cyr} - \log z_{cy}\Big) + \frac{\log e_{cr}}{\lambda^2} - \frac{\log z_{c+}}{\lambda^2} - \frac{\alpha_r}{\lambda^2} + \frac{\beta_0}{\sigma_\beta^2}$$

$$\Delta_{cr}^{-1} = \frac{y}{\tau^2} + \frac{1}{\lambda^2} + \frac{1}{\sigma_\beta^2}$$

$$\theta_{cr} = \Delta_{cr} \times \Delta_{cr}^{-1}\theta_{cr}$$

α **posterior**

$$p(\boldsymbol{\alpha} \mid \boldsymbol{e}, \boldsymbol{\beta}, \lambda^2, \boldsymbol{z}) \propto \prod_{cr} \exp\Big[-\frac{1}{2\lambda^2}\Big(\log e_{cr} - (\log z_{c+} + \beta_{cr} + \alpha_r)\Big)^2\Big] \cdot \prod_r \exp\Big[-\frac{1}{2\sigma_\alpha^2}\Big(\alpha_r - \alpha_0\Big)^2\Big]$$

$$\propto \prod_{cr} \exp\Big[-\frac{1}{2\lambda^2}\Big((\log e_{cr})^2 + (\log z_{c+})^2 + \beta_{cr}^2 + \alpha_r^2 - 2\log e_{cr}\log z_{c+} - 2\beta_{cr}\log e_{cr} - 2\alpha_r \log e_{cr} + 2\beta_{cr}\log z_{c+} + 2\alpha_r \log z_{c+} + 2\beta_{cr}\alpha_r\Big)\Big] \cdot \prod_r \exp\Big[-\frac{1}{2\sigma_\alpha^2}\Big(\alpha_r^2 + \alpha_0^2 - 2\alpha_r\alpha_0\Big)\Big]$$

$$\propto \prod_r \exp\Big[-\frac{1}{2}\Big(\alpha_r^2\Big\{\frac{c}{\lambda^2} + \frac{1}{\sigma_\alpha^2}\Big\} - 2\alpha_r\Big\{\sum_c \Big(\frac{\log e_{cr}}{\lambda^2} - \frac{\log z_{c+}}{\lambda^2} - \frac{\beta_{cr}}{\lambda^2}\Big) + \frac{\alpha_0}{\sigma_\alpha^2}\Big\}\Big)\Big]$$

$$\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{e}, \lambda^2 \sim N(\boldsymbol{\phi}, \boldsymbol{\Omega})$$

where

$$\Omega_r^{-1}\phi_r = \frac{1}{\lambda^2}\sum_c \Big(\log e_{cr} - \log z_{c+} - \beta_{cr}\Big) + \frac{\alpha_0}{\sigma_\alpha^2}$$

$$\Omega_r^{-1} = \frac{c}{\lambda^2} + \frac{1}{\sigma_\alpha^2}$$

$$\phi_r = \Omega_r \times \Omega_r^{-1}\phi_r$$

τ^2 posterior

$$p(\tau^2 \mid \boldsymbol{x}, \boldsymbol{\beta}, \tau_a, \tau_b, \boldsymbol{z}) \propto \Big\{ \Big(\frac{1}{2\pi\tau^2} \Big)^{\frac{CYR}{2}} \prod_{cyr} \exp \Big[- \frac{1}{2\tau^2} \Big(\log x_{cyr} - (\log z_{cy} + \beta_{cr}) \Big)^2 \Big] \Big\} \cdot \Big(\frac{1}{\tau^2} \Big)^{\tau_a + 1} \exp \Big[- \frac{\tau_a}{\tau^2} \Big]$$
$$\propto \Big(\frac{1}{\tau^2} \Big)^{\frac{CYR}{2} + \tau_a + 1} \exp \Big[- \frac{1}{\tau^2} \Big(\frac{1}{2} \sum_{cyr} \Big\{ \log x_{cyr} - (\log z_{cy} + \beta_{cr}) \Big\}^2 + \tau_b \Big) \Big]$$

$$\tau^2 \mid \boldsymbol{x}, \boldsymbol{\beta}, \tau_a, \tau_b, \boldsymbol{z} \sim IG\Big(\frac{CYR}{2} + \tau_a, \frac{1}{2} \sum_{cyr} \Big\{ \log x_{cyr} - (\log z_{cy} + \beta_{cy}) \Big\}^2 + \tau_b \Big)$$

λ^2 posterior

$$p(\lambda^2 \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda_a, \lambda_b, \boldsymbol{z}, \boldsymbol{e}) \propto \Big\{ \Big(\frac{1}{2\pi\lambda^2} \Big)^{\frac{CR}{2}} \prod_{cr} \exp \Big[- \frac{1}{2\lambda^2} \Big(\log e_{cr} - (\log z_{c+} + \alpha_r + \beta_{cr}) \Big)^2 \Big] \Big\} \cdot \Big(\frac{1}{\lambda^2} \Big)^{\lambda_a + 1} \exp \Big[- \frac{\lambda_a}{\lambda^2} \Big]$$
$$\propto \Big(\frac{1}{\lambda^2} \Big)^{\frac{CR}{2} + \lambda_a + 1} \exp \Big[- \frac{1}{\lambda^2} \Big(\frac{1}{2} \sum_{cr} \Big\{ \log e_{cr} - (\log z_{c+} + \alpha_r + \beta_{cr}) \Big\}^2 + \lambda_b \Big) \Big]$$

$$\lambda^2 \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda_a, \lambda_b, \boldsymbol{z}, \boldsymbol{e} \sim IG\Big(\frac{CR}{2} + \lambda_a, \frac{1}{2} \sum_{cr} \Big\{ \log e_{cr} - (\log z_{c+} + \alpha_r + \beta_{cr}) \Big\}^2 + \lambda_b \Big)$$

B.5 Posterior models for car age exposure modelling

$$
\begin{aligned}
\log \boldsymbol{a} &\sim N(\boldsymbol{\theta}, \delta^2) \\
\boldsymbol{\theta} &\sim N(\boldsymbol{\theta_0}, \sigma_\theta^2) \text{ where } \sigma_\theta^2 \text{ large} \\
\delta^2 &\sim IG(\delta_a, \delta_b)
\end{aligned}
$$

$$
t_{cygr} = \frac{\theta_{cyg}}{\sum_g \theta_{cyg}} \times x_{cyr}
$$

$$
\begin{aligned}
p(\boldsymbol{\theta}, \delta^2 \mid \log \boldsymbol{a}) &\propto p(\log \boldsymbol{a} \mid \boldsymbol{\theta}, \delta^2) p(\boldsymbol{\theta}) p(\delta^2) \\
&\propto \Big(\frac{1}{2\pi\delta^2}\Big)^{CYG/2} \exp \sum_{cyg} \Big(- \frac{1}{2\delta^2} (\log a_{cyg} - \theta_{cyg})^2 \Big) \\
&\cdot \Big(\frac{1}{2\pi\sigma_\theta^2}\Big)^{CYG/2} \exp \sum_{cyg} \Big(- \frac{1}{2\sigma_\theta^2} (\theta_{cyg} - \theta_0)^2 \Big) \\
&\cdot \frac{1}{\delta^2}^{\delta_a + 1} \exp \Big(- \frac{\delta_b}{\delta^2} \Big)
\end{aligned}
$$

Posterior for θ

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \delta^2, \log \boldsymbol{a}) &\propto \exp \sum_{cyg} \Big(- \frac{1}{2\delta^2} (\log a_{cyg} - \theta_{cyg})^2 - \frac{1}{2\sigma_\theta^2} (\theta_{cyg} - \theta_0)^2 \Big) \\
&\propto \exp \sum_{cyg} \Big(- \frac{1}{2} \Big[\frac{(\log a_{cyg})^2}{\delta^2} + \frac{\theta_{cyg}^2}{\delta^2} - \frac{2\theta_{cyg} \log a_{cyg}}{\delta^2} + \frac{\theta_{cyg}^2}{\sigma_\theta^2} + \frac{\theta_0^2}{\sigma_\theta^2} - \frac{2\theta_{cyg}\theta_0}{\sigma_\theta^2} \Big] \\
&\propto \exp \sum_{cyg} \Big(- \frac{1}{2} \Big[\theta_{cyg}^2 \Big\{ \frac{1}{\delta^2} + \frac{1}{\sigma_\theta^2} \Big\} - 2\theta_{cyg} \Big\{ \frac{\log a_{cyg}}{\delta^2} + \frac{\theta_0}{\sigma_\theta^2} \Big\} + \dots \Big] \Big)
\end{aligned}
$$

$$
\boldsymbol{\theta} \sim N\Big(\frac{\sigma_\theta^2 \log a_{cyg} + \delta^2 \theta_0}{\delta^2 + \sigma_\theta^2}, \frac{\delta^2 \sigma_\theta^2}{\delta^2 + \sigma_\theta^2} \Big)
$$

Posterior for δ

$$
\begin{aligned}
p(\delta^2 \mid \boldsymbol{\theta}, \log \boldsymbol{a}) &\propto \Big(\frac{1}{2\pi\delta^2}\Big)^{CYG/2} \exp \sum_{cyg} \Big(-\frac{1}{2\delta^2}(\log a_{cyg} - \theta_{cyg})^2 \Big) \\
&\quad \cdot \frac{1}{\delta^2}^{\delta_a+1} \exp\Big(-\frac{\delta_b}{\delta^2}\Big) \\
&\propto \Big(\frac{1}{2\pi}\Big)^{\frac{CYG}{2}} \Big(\frac{1}{\delta^2}\Big)^{\delta_a+1+\frac{cyg}{2}} \exp -\frac{1}{\delta^2}\Big(\frac{1}{2}\sum_{cyg}\big(\log a_{cyg} - \theta\big)^2 + \delta_b\Big)
\end{aligned}
$$

$$
\delta^2 \sim IG\Big(\frac{CYG}{2} + \delta_a, \frac{1}{2}\sum_{cyg}\big(\log a_{cyg} - \theta_{cyg})^2 + \delta_b\Big)
$$

Appendix C

Results from large exposure modelling - prior 2 and 3

Table C.1: Prior values for log-Normal exposure model on 12 year dataset

	Prior 0	Prior 1	Prior 2	Prior 3
β_0	0	0	0	0
σ_β	1000	1000	1000	1000
α_0	0	0	0	0
σ_α	1000	1000	1000	1000
τ_a	5.0	5.0	5.0	5.0
τ_b	0.5	0.06	0.01	0.003
λ_a	3	3	3	3
λ_b	0.25	0.1	0.025	0.007

C.1 Modelling results

C.1.1 Prior 2

Figures C.1, C.2 and C.3 show the modelled values of x_{cyr} that result from the simulation with Prior 2, for Motorways, A roads and Minor roads respectively.

These values represent an estimate of the log of the number of billion vehicle kilometres travelled by each car type in each year on each road type, which we call disaggregated traffic data.

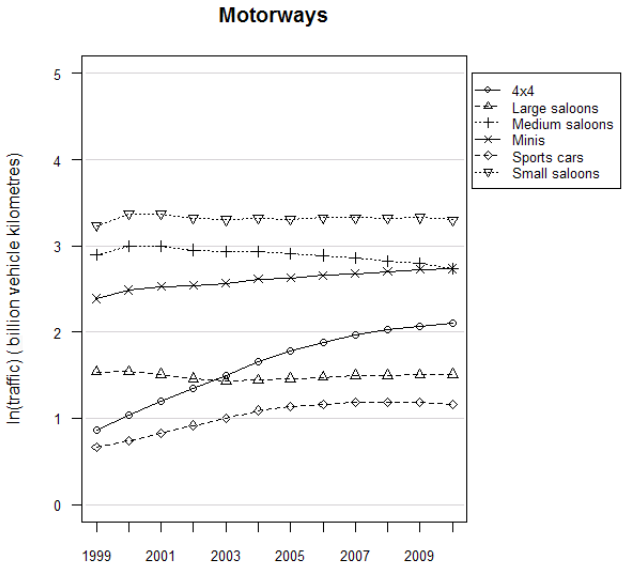


Figure C.1: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on Motorways on log-Normal exposure model with less diffuse prior (Prior 2)

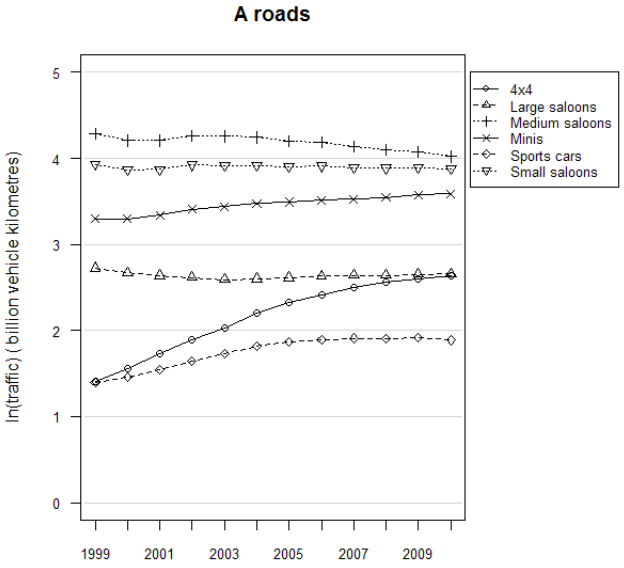


Figure C.2: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on A roads on log-Normal exposure model with less diffuse prior (Prior 2)

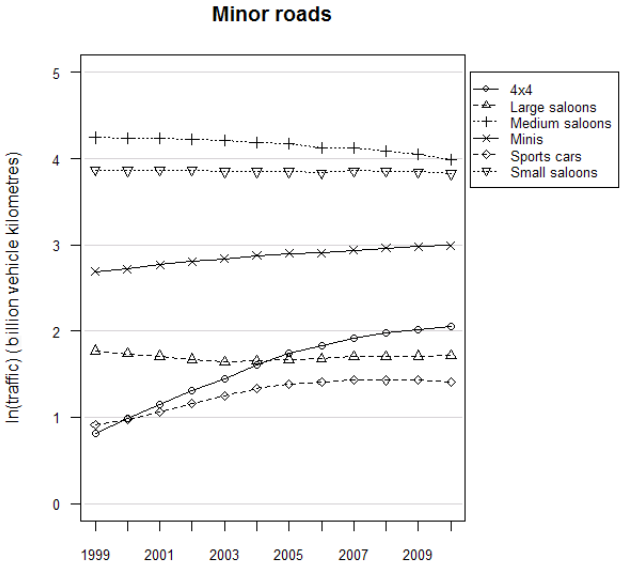


Figure C.3: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on Minor roads on log-Normal exposure model with less diffuse prior (Prior 2)

C.1.2 Prior 3

Figures C.4, C.5 and C.6 show the modelled values of x_{cyr} that result from the simulation with Prior 3, for Motorways, A roads and Minor roads respectively. These values represent an estimate of the log of the number of billion vehicle kilometres travelled by each car type in each year on each road type, which we call disaggregated traffic data.

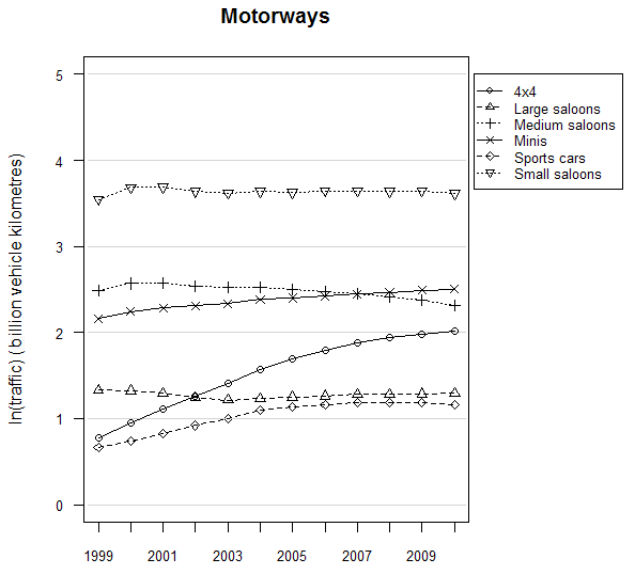


Figure C.4: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on Motorways on log-Normal exposure model with precise prior (Prior 3)

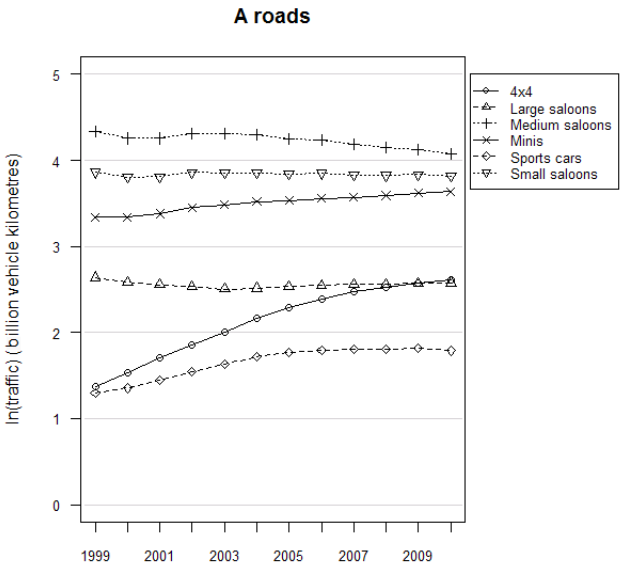


Figure C.5: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on A roads on log-Normal exposure model with precise prior (Prior 3)

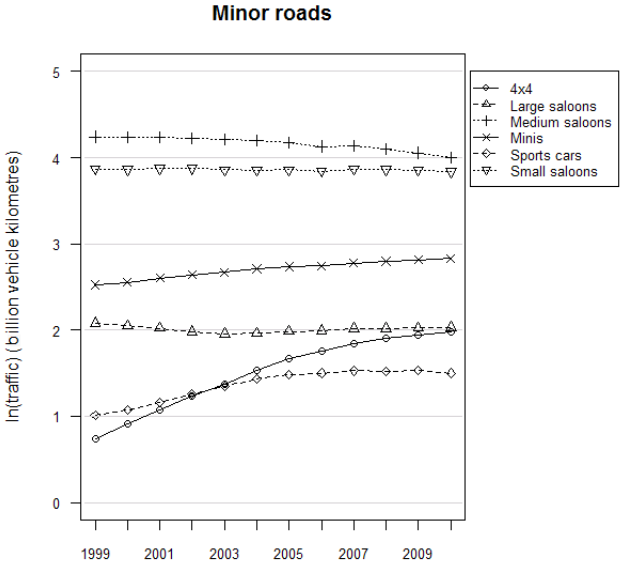


Figure C.6: Modelled disaggregated exposure (traffic flow) x_{cyr} by year and car type on Minor roads on log-Normal exposure model with precise prior (Prior 3)

C.2 Multivariate t-testing

C.2.1 Prior 2

Correlations

For a range of k, the correlation between the exact and test quantiles (in the QQ-plot) are calculated. Figure C.7 displays the range of correlation values from $k = 15$ to $k = 100$ along with a 95% confidence envelope generated from randomly simulated F distributed variables. The highest correlation is 0.9994 which occurs when $k = 45$, and correlations throughout the pictured range are within the simulated 95% confidence intervals.

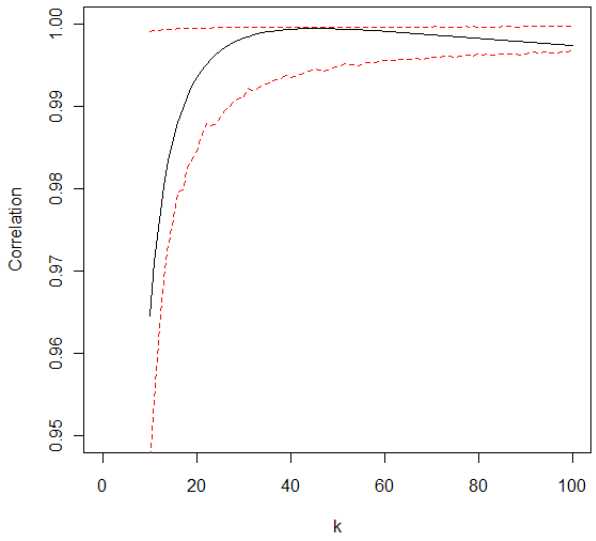


Figure C.7: Plot of test correlations for testing MVt_k of exposure model posterior distribution over range of k – less diffuse prior (Prior 2)

Variance about the QQ-line

Figure C.8 shows the range of absolute differences between the exact and data quantiles for $k = 20, \ldots, 200$ and the associated acceptability bands. The degrees

of freedom k with the smallest sum is $k = 60$, with a difference sum of 4.2. Comparing this to a set of randomly generated quantiles suggests that differences from $k = 45$ to $k = 75$ are approximately within a simulated 95% acceptability range.

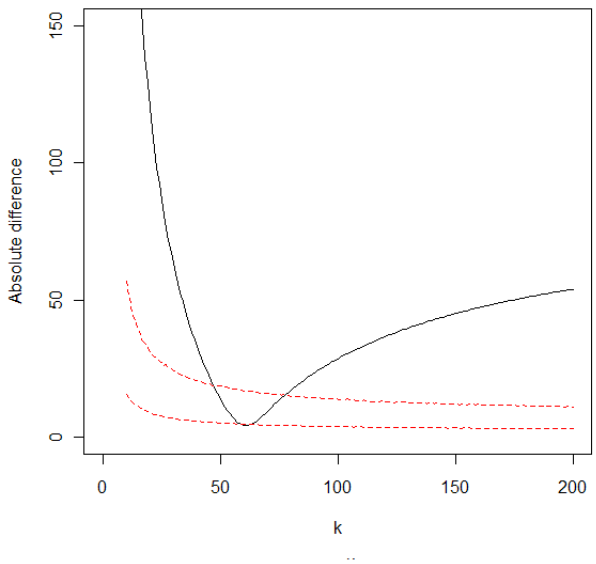


Figure C.8: Plot of absolute variance from QQ-plot for testing MVt_k of exposure model posterior distribution over range of k – less diffuse prior (Prior 2)

For the purposes of the modelling in Chapter 5 we have selected a value of $k = 50$.

C.2.2 Prior 3

Correlations

For a range of k, the correlation between the exact and test quantiles (in the QQ-plot) are calculated. Figure C.9 displays the range of correlation values from $k = 15$ to $k = 100$ along with a 95% confidence envelope generated from randomly simulated F distributed variables. The highest correlation is 0.9990 which occurs when $k = 48$, and correlations throughout the pictured range are within the

simulated 95% confidence intervals.

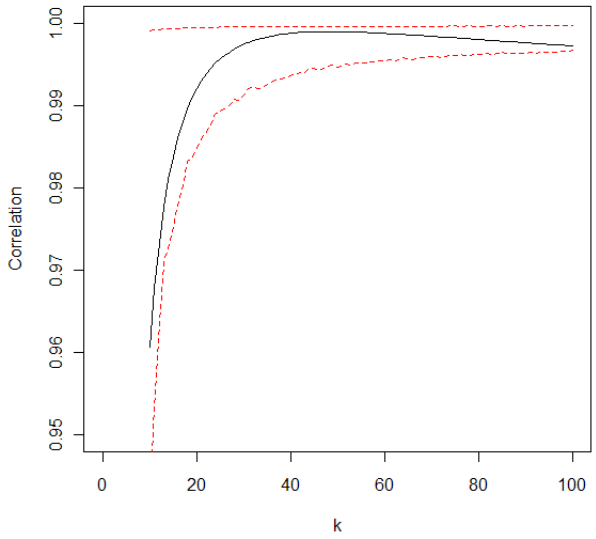


Figure C.9: Plot of test correlations for testing MVt_k of exposure model posterior distribution over range of k – precise prior (Prior 3)

Variance about the QQ-line

Figure C.10 shows the range of absolute differences between the exact and data quantiles for $k = 20, \ldots, 200$ and the associated acceptability bands. The degrees of freedom k with the smallest sum is $k = 45$, with a difference sum of 7.0. Comparing this to a set of randomly generated quantiles suggests that differences from $k = 35$ to $k = 55$ are approximately within a simulated 95% acceptability range.

For the purposes of the modelling in Chapter 5 we have selected a value of $k = 50$.

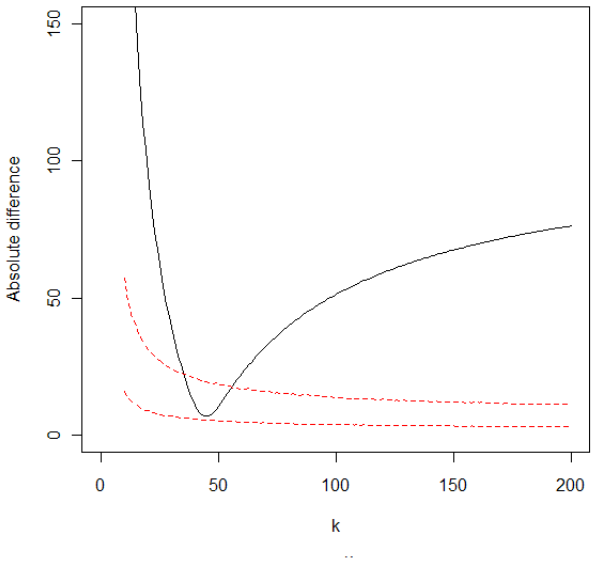


Figure C.10: Plot of absolute variance from QQ-plot for testing MVt_k of exposure model posterior distribution over range of k – precise prior (Prior 3)

Appendix D

Accident rate modelling tables

Table D.1: Marginal likelihoods and model probabilities for accident rate model

	Model	Marginal likelihood	Model choice	Model choice by group	BIC
1	ME	39462.8	0%	100%	17394.9
ME					
2	+cy	39453.5	0%	0%	17384.0
3	+cg	39603.9	0%	0%	17120.5
4	+cr	40090.5	0%	100%	13680.0
5	+cb	39482.3	0%	0%	17347.2
6	+yg	39481.8	0%	0%	17336.3
7	+yr	39465.2	0%	0%	17374.4
8	+yb	22627.5	0%	0%	17390.8
9	+gr	39464.6	0%	0%	17079.2
10	+gb	39454.6	0%	0%	17408.7
11	+rb	39951.7	0%	0%	16413.1
ME + cr					
12	+cy	40091.1	0%	0%	13668.5
13	+cg	40233.9	0%	0%	13405.5
14	+cb	40116.9	0%	0%	13632.3
15	+yg	40112.8	0%	0%	13621.4
16	+yr	40090.8	0%	0%	13678.0
17	+yb	40091.7	0%	0%	13675.9
18	+gr	40097.3	0%	0%	13311.2
19	+gb	40085.7	0%	0%	13693.8
20	+rb	40584.4	0%	100%	12698.2

Table D.1 – continued from previous page

ME + cr + rb					
21	+cy	40579.3	0%	0%	12686.6
22	+cg	40725.2	0%	100%	12423.6
23	+cb	40587.8	0%	0%	12690.7
24	+yg	40603.3	0%	0%	12639.5
25	+yr	40579.5	0%	0%	12696.1
26	+yb	40582.4	0%	0%	12694.7
27	+gr	40588.5	0%	0%	12329.3
28	+gb	40575.6	0%	0%	12712.5
ME + cr + rb + cg					
29	+cy	40719.5	0%	0%	12412.8
30	+cb	40728.1	0%	0%	12416.1
31	+yg	40745.7	0%	100%	12364.7
32	+yr	40720.9	0%	0%	12421.5
33	+yb	40723.1	0%	0%	12420.1
34	+gr	40730.1	0%	0%	12048.3
35	+gb	40715.7	0%	0%	12437.7
ME + cr + rb + cg + yg					
36	+cy	40738.8	0%	0%	12353.8
37	+cb	40746.5	0%	5%	12357.2
38	+yr	40744.6	0%	1%	12362.6
39	+yb	40748.2	0%	26%	12361.3
40	+gr	40749.1	1%	68%	11989.6
41	+gb	40743.4	0%	0%	12378.8
ME + cr + rb + cg + yg + gr					
42	+cy	40747.1	0%	0%	11978.6
43	+cb	40753.7	48%	50%	11982.1
44	+yr	40746.6	0%	0%	11983.1
45	+yb	40753.7	49%	50%	11986.1
46	+gb	40744.0	0%	0%	12015.6
ME + cr + rb + cg + yg + gr + yb					
47	+cy	40743.5	0%	0%	11974.7
48	+cb	40750.5	2%	100%	11979.5
49	+yr	40743.6	0%	0%	11980.9
50	+gb	40740.8	0%	0%	12011.0
ME + cr + rb + cg + yg + gr + yb + cb					
51	+cy	40746.6	0%	75%	11967.0
52	+yr	40743.0	0%	2%	11974.3
53	+gb	40745.4	0%	23%	12004.7
ME + cr + rb + cg + yg + gr + yb + cb + cy					
54	+yr	40745.9	0%	100%	11960.2

Table D.1 – continued from previous page

55	+gb	40739.4	0%	0%	11992.1
ME + cr + rb + cg + yg + gr + yb + cb + cy + yr					
56	+gb	40738.8	0%	100%	11985.4

Table D.2: Mean and standard deviation of coefficients for high probability accident rate models with fixed exposure

		Model 43		Model 45	
		Mean	SD	Mean	SD
Constant		1.61	0.08	1.57	0.08
Car size	4x4	-	-	-	-
	Large	-0.42	0.11	-0.42	0.11
	Medium	-0.70	0.09	-0.68	0.09
	Small	-0.45	0.08	-0.42	0.08
	Minis	0.03	0.08	0.07	0.08
	Sports	0.08	0.11	0.13	0.11
Year		-0.07	0.00	-0.07	0.00
Car age	0-2	-	-	-	-
	3-5	0.04	0.09	0.03	0.09
	6-10	-0.16	0.09	-0.17	0.08
	11-15	0.17	0.10	0.16	0.09
	16+	1.19	0.14	1.18	0.13
Road	M	-	-	-	-
	A	-0.03	0.08	-0.07	0.08
	Minor	0.28	0.08	0.22	0.08
Bend		-1.52	0.06	-1.30	0.05
Car size	4x4 & 0-2	-	-	-	-
& car age	4x4 & 3-5	-	-	-	-
	4x4 & 6-10	-	-	-	-
	4x4 & 11-15	-	-	-	-
	4x4 & 16+	-	-	-	-
	large & 0-2	-	-	-	-
	large & 3-5	0.04	0.11	0.05	0.11
	large & 6-10	0.14	0.10	0.14	0.10
	large & 11-15	0.08	0.11	0.08	0.11
	large & 16+	-0.81	0.14	-0.81	0.14
	medium & 0-2	-	-	-	-
	medium & 3-5	0.07	0.08	0.07	0.08
	medium & 6-10	0.18	0.08	0.18	0.07
	medium & 11-15	0.26	0.09	0.27	0.09

Table D.2 – continued from previous page

	medium & 16+	-0.39	0.12	-0.38	0.12
	small & 0-2	-	-	-	-
	small & 3-5	-0.05	0.08	-0.05	0.07
	small & 6-10	0.12	0.07	0.12	0.07
	small & 11-15	0.22	0.08	0.22	0.08
	small & 16+	-0.44	0.11	-0.44	0.11
	mini & 0-2	-	-	-	-
	mini & 3-5	-0.12	0.07	-0.11	0.07
	mini & 6-10	-0.08	0.07	-0.08	0.07
	mini & 11-15	-0.08	0.08	-0.08	0.08
	mini & 16+	-0.88	0.11	-0.88	0.11
	sport & 0-2	-	-	-	-
	sport & 3-5	-0.10	0.09	-0.09	0.09
	sport & 6-10	-0.11	0.09	-0.11	0.09
	sport & 11-15	-0.18	0.10	-0.18	0.10
	sport & 16+	-1.74	0.13	-1.73	0.13
Car size	4x4 & M	-	-	-	-
& road	4x4 & A	-	-	-	-
	4x4 & min	-	-	-	-
	large & M	-	-	-	-
	large & A	-0.39	0.10	-0.37	0.11
	large & min	0.18	0.11	0.20	0.11
	medium & M	-	-	-	-
	medium & A	0.06	0.09	0.09	0.09
	medium & min	-0.57	0.09	-0.53	0.09
	small & M	-	-	-	-
	small & A	0.50	0.08	0.54	0.08
	small & min	0.07	0.08	0.12	0.08
	mini & M	-	-	-	-
	mini & A	0.59	0.08	0.65	0.08
	mini & min	0.63	0.08	0.70	0.08
	sport & M	-	-	-	-
	sport & A	0.64	0.11	0.70	0.11
	sport & min	0.17	0.11	0.24	0.11
Car size	4x4 & bend	-	-		
& bend	large & bend	0.06	0.07		
	medium & bend	0.16	0.06		
	small & bend	0.18	0.05		
	mini & bend	0.24	0.05		
	sport & bend	0.25	0.07		
Year &	0-2	-	-	-	-
car age	3-5	0.01	0.01	0.01	0.01

Table D.2 – continued from previous page

	6-10	0.03	0.01	0.03	0.01
	11-15	0.00	0.01	0.00	0.01
	16+	0.01	0.01	0.01	0.01
Year &	bend			-0.01	0.00
Car age	0-2 & M	-	-	-	-
& road	0-2 & A	-	-	-	-
	0-2 & min	-	-	-	-
	3-5 & M	-	-	-	-
	3-5 & A	0.15	0.06	0.16	0.06
	3-5 & min	0.21	0.06	0.22	0.06
	6-10 & M	-	-	-	-
	6-10 & A	0.51	0.06	0.51	0.06
	6-10 & min	0.62	0.06	0.62	0.06
	11-15 & M	-	-	-	-
	11-15 & A	0.69	0.07	0.69	0.07
	11-15 & min	0.98	0.07	0.99	0.07
	16+ & M	-	-	-	-
	16+ & A	0.24	0.10	0.25	0.10
	16+ & min	0.67	0.10	0.67	0.10
	M & bend	-	-	-	-
	A & bend	0.93	0.05	0.94	0.05
	min & bend	1.23	0.05	1.24	0.05
σ^2_ϕ		59.90	7.48	57.66	6.77
Deviance		10855.82	51.24	10855.89	50.51

Table D.3: Mean and standard deviation of coefficients for high probability accident rate models with variable exposure

		Model 43		Model 45	
		Mean	SD	Mean	SD
Constant		1.64	0.40	1.60	0.39
Car size	4x4	-	-	-	-
	Large	-0.30	0.67	-0.29	0.67
	Medium	-0.73	0.39	-0.71	0.39
	Small	-0.46	0.54	-0.43	0.54
	Minis	0.05	0.64	0.09	0.64
	Sports	0.15	0.64	0.20	0.64
Year		-0.07	0.01	-0.07	0.01
Car age	0-2	-	-	-	-
	3-5	0.07	0.13	0.07	0.13

Table D.3 – continued from previous page

	6-10	-0.13	0.13	-0.13	0.13
	11-15	0.21	0.15	0.21	0.15
	16+	1.21	0.16	1.21	0.16
Road	M	-	-	-	-
	A	0.01	0.50	-0.02	0.50
	Minor	0.27	0.51	0.23	0.52
Bend		-1.49	0.13	-1.29	0.09
Car size	4x4 & 0-2	-	-	-	-
& car age	4x4 & 3-5	-	-	-	-
	4x4 & 6-10	-	-	-	-
	4x4 & 11-15	-	-	-	-
	4x4 & 16+	-	-	-	-
	large & 0-2	-	-	-	-
	large & 3-5	0.03	0.13	0.03	0.13
	large & 6-10	0.12	0.13	0.13	0.12
	large & 11-15	0.06	0.13	0.07	0.12
	large & 16+	-0.82	0.16	-0.81	0.16
	medium & 0-2	-	-	-	-
	medium & 3-5	0.06	0.11	0.06	0.11
	medium & 6-10	0.17	0.10	0.18	0.10
	medium & 11-15	0.25	0.11	0.26	0.10
	medium & 16+	-0.39	0.14	-0.39	0.13
	small & 0-2	-	-	-	-
	small & 3-5	-0.07	0.10	-0.06	0.09
	small & 6-10	0.11	0.09	0.11	0.09
	small & 11-15	0.20	0.10	0.20	0.10
	small & 16+	-0.45	0.13	-0.45	0.12
	mini & 0-2	-	-	-	-
	mini & 3-5	-0.13	0.10	-0.13	0.10
	mini & 6-10	-0.10	0.10	-0.09	0.09
	mini & 11-15	-0.11	0.10	-0.10	0.10
	mini & 16+	-0.89	0.13	-0.89	0.13
	sport & 0-2	-	-	-	-
	sport & 3-5	-0.11	0.12	-0.10	0.11
	sport & 6-10	-0.12	0.11	-0.12	0.11
	sport & 11-15	-0.20	0.12	-0.19	0.12
	sport & 16+	-1.74	0.15	-1.74	0.15
Car size	4x4 & M	-	-	-	-
& road	4x4 & A	-	-	-	-
	4x4 & min	-	-	-	-
	large & M	-	-	-	-
	large & A	-0.47	0.87	-0.46	0.87

Table D.3 – continued from previous page

	large & min	0.25	0.86	0.26	0.86
	medium & M	-	-	-	-
	medium & A	0.06	0.56	0.09	0.56
	medium & min	-0.49	0.58	-0.46	0.58
	small & M	-	-	-	-
	small & A	0.51	0.81	0.55	0.80
	small & min	0.07	0.80	0.11	0.80
	mini & M	-	-	-	-
	mini & A	0.54	0.79	0.58	0.78
	mini & min	0.68	0.80	0.73	0.80
	sport & M	-	-	-	-
	sport & A	0.59	0.87	0.64	0.87
	sport & min	0.18	0.84	0.24	0.84
Car size	4x4 & bend	-	-		
& bend	large & bend	0.05	0.09		
	medium & bend	0.15	0.08		
	mini & bend	0.23	0.08		
	sport & bend	0.24	0.08		
	small & bend	0.17	0.07		
Year &	0-2	-	-	-	-
car age	3-5	0.01	0.01	0.01	0.01
	6-10	0.03	0.01	0.03	0.01
	11-15	0.00	0.01	0.00	0.01
	16+	0.01	0.01	0.01	0.01
Year &	bend	-		-0.01	0.00
Car age	0-2 & M	-	-	-	-
& road	0-2 & A	-	-	-	-
	0-2 & min	-	-	-	-
	3-5 & M	-	-	-	-
	3-5 & A	0.13	0.09	0.14	0.10
	3-5 & min	0.19	0.09	0.19	0.09
	6-10 & M	-	-	-	-
	6-10 & A	0.49	0.10	0.49	0.10
	6-10 & min	0.59	0.09	0.59	0.10
	11-15 & M	-	-	-	-
	11-15 & A	0.66	0.11	0.67	0.11
	11-15 & min	0.95	0.11	0.95	0.11
	16+ & M	-	-	-	-
	16+ & A	0.22	0.12	0.22	0.12
	16+ & min	0.65	0.12	0.65	0.12
	M & bend	-	-	-	-
	A & bend	0.91	0.09	0.92	0.09

Table D.3 – continued from previous page

	min & bend	1.21	0.09	1.22	0.08
σ^2_ϕ		41.92	7.44	40.86	6.98

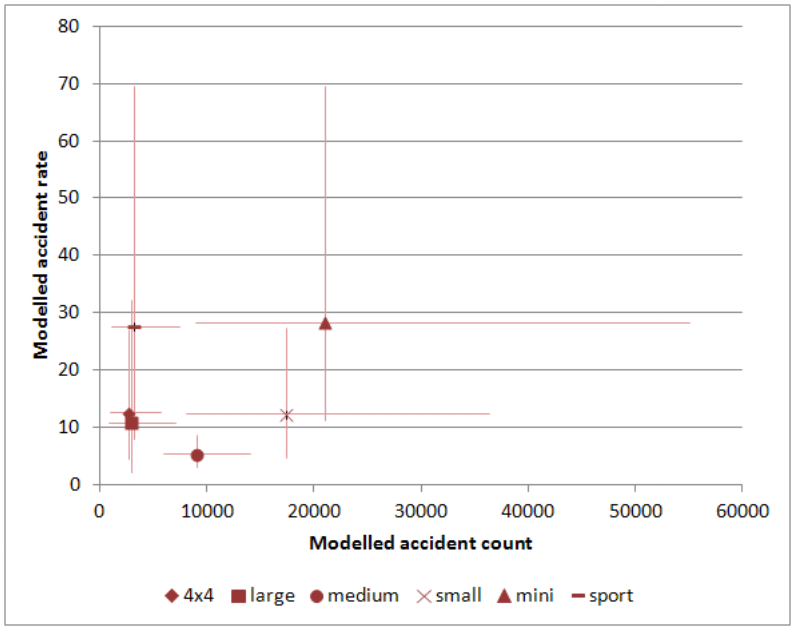


Figure D.1: Relationship between modelled accident rate and accident propensity for car type with associated posterior intervals

Appendix E

Accident severity modelling tables

Table E.1: Marginal likelihoods and model probabilities for accident severity model

	Model	Marginal likelihood	Model choice	Model choice by group	BIC
1	ME	-17032.3	0%	0%	9236.9
ME					
2	+cy	-17050.4	0%	0%	9688.3
3	+cg	-17069.1	0%	0%	9433.2
4	+cr	-17051.6	0%	0%	9358.1
5	+co	-17034.9	0%	0%	9304.5
6	+cb	-17041.5	0%	0%	9317.3
7	+yg	-17046.1	0%	0%	9612.3
8	+yr	-17038.2	0%	0%	9464.1
9	+yo	-17032.6	0%	0%	9386.8
10	+yb	-17032.4	0%	0%	9386.4
11	+gr	-17025.6	16%	61%	9331.5
12	+go	-17035.7	0%	0%	9296.9
13	+gb	-17040.1	0%	0%	9305.3
14	+ro	-17033.2	0%	0%	9276.0
15	+rb	-17030.5	0%	0%	9270.7
16	+ob	-17026.0	10%	39%	9253.4
ME + gr					

Table E.1 – continued from previous page

17	+cy	-17048.6	0%	0%	9782.9
18	+cg	-17067.2	0%	0%	9529.3
19	+cr	-17049.8	0%	0%	9452.8
20	+co	-17033.1	0%	0%	9362.3
21	+cb	-17039.6	0%	0%	9374.6
22	+yg	-17044.3	0%	0%	9706.8
23	+yr	-17036.3	0%	0%	9558.7
24	+yo	-17030.8	0%	0%	9407.1
25	+yb	-17030.6	0%	0%	9406.6
26	+go	-17033.9	0%	0%	9391.2
27	+gb	-17038.2	0%	0%	9369.7
28	+ro	-17031.4	0%	0%	9352.1
29	+rb	-17028.7	1%	1%	9347.0
30	+ob	-17024.2	65%	99%	9335.6
ME + gr + ob					
31	+cy	-17047.1	0%	0%	9786.9
32	+cg	-17065.8	0%	0%	9533.4
33	+cr	-17048.4	0%	0%	9456.9
34	+co	-17031.9	0%	0%	9404.1
35	+cb	-17038.4	0%	0%	9416.3
36	+yg	-17042.8	0%	0%	9710.8
37	+yr	-17034.9	0%	0%	9562.9
38	+yo	-17029.4	0%	5%	9485.8
39	+yb	-17029.0	1%	8%	9484.7
40	+go	-17032.5	0%	0%	9395.5
41	+gb	-17036.6	0%	0%	9404.7
42	+ro	-17030.2	0%	2%	9375.4
43	+rb	-17026.6	6%	85%	9349.7
ME + gr + ob + rb					
44	+cy	-17049.5	0%	0%	9801.0
45	+cg	-17068.3	0%	0%	9547.6
46	+cr	-17051.0	0%	0%	9471.4
47	+co	-17034.5	0%	2%	9418.6
48	+cb	-17040.9	0%	0%	9430.7
49	+yg	-17045.2	0%	0%	9725.0
50	+yr	-17037.3	0%	0%	9576.9
51	+yo	-17031.8	0%	30%	9499.9
52	+yb	-17031.3	0%	50%	9498.7
53	+go	-17035.0	0%	1%	9409.6
54	+gb	-17039.2	0%	0%	9387.9
55	+ro	-17032.5	0%	16%	9389.2
ME + gr + ob + rb + yb					

Table E.1 – continued from previous page

56	+cy	-17054.3	0%	0%	9875.6
57	+cg	-17073.1	0%	0%	9622.2
58	+cr	-17055.8	0%	0%	9546.1
59	+co	-17039.3	0%	4%	9493.3
60	+cb	-17045.6	0%	0%	9467.8
61	+yg	-17049.9	0%	0%	9799.4
62	+yr	-17042.0	0%	0%	9651.5
63	+yo	-17036.6	0%	60%	9574.5
64	+go	-17039.7	0%	3%	9484.1
65	+gb	-17044.0	0%	0%	9462.5
66	+ro	-17037.2	0%	33%	9463.8
ME + gr + ob + rb + yb + yo					
67	+cy	-17059.4	0%	0%	10025.6
68	+cg	-17078.4	0%	0%	9772.5
69	+cr	-17061.1	0%	0%	9696.3
70	+co	-17044.2	0%	15%	9642.8
71	+cb	-17050.9	0%	0%	9655.2
72	+yg	-17055.3	0%	0%	9949.7
73	+yr	-17047.3	0%	1%	9801.8
74	+go	-17045.1	0%	6%	9634.6
75	+gb	-17049.3	0%	0%	9643.7
76	+ro	-17042.6	0%	79%	9614.1
ME + gr + ob + rb + yb + yo + ro					
77	+cy	-17065.4	0%	0%	10065.1
78	+cg	-17084.4	0%	0%	9812.1
79	+cr	-17067.0	0%	0%	9735.9
80	+co	-17050.1	0%	72%	9682.3
81	+cb	-17056.8	0%	0%	9694.7
82	+yg	-17061.2	0%	0%	9989.3
83	+yr	-17053.3	0%	3%	9841.3
84	+go	-17051.2	0%	25%	9674.4
85	+gb	-17055.2	0%	0%	9683.3
ME + gr + ob + rb + yb + yo + ro + co					
86	+cy	-17073.2	0%	0%	10133.9
87	+cg	-17091.3	0%	0%	9879.3
88	+cr	-17074.7	0%	0%	9804.3
89	+cb	-17064.3	0%	0%	9725.7
90	+yg	-17068.8	0%	0%	10057.6
91	+yr	-17060.8	0%	9%	9009.5
92	+go	-17058.5	0%	89%	9742.1
93	+gb	-17062.8	0%	1%	9751.5
ME + gr + ob + rb + yb + yo + ro + co + go					

Table E.1 – continued from previous page

94	+cy	-17081.6	0%	0%	10193.8
95	+cg	-17099.8	0%	0%	9939.2
96	+cr	-17083.2	0%	0%	9864.0
97	+cb	-17072.7	0%	3%	9785.5
98	+yg	-17077.2	0%	0%	10117.5
99	+yr	-17069.2	0%	86%	9969.3
100	+gb	-17071.3	0%	11%	9780.5
ME + gr + ob + rb + yb + yo + ro + co + go + yr					
101	+cy	-17092.3	0%	0%	10421.0
102	+cg	-17110.5	0%	0%	10166.4
103	+cr	-17093.9	0%	0%	10091.3
104	+cb	-17083.4	0%	19%	10012.7
105	+yg	-17087.9	0%	0%	10344.6
106	+gb	-17081.9	0%	80%	10007.7
ME + gr + ob + rb + yb + yo + ro + co + go + yr + gb					
107	+cy	-17105.1	0%	0%	10459.3
108	+cg	-17123.3	0%	0%	10204.8
109	+cr	-17106.6	0%	0%	10129.6
110	+cb	-17096.0	0%	99%	10050.9
111	+yg	-17100.6	0%	1%	10383.0
ME + gr + ob + rb + yb + yo + ro + co + go + yr + gb + cb					
112	+cy	-17119.1	0%	1%	10502.5
113	+cg	-17137.2	0%	0%	10247.8
114	+cr	-17120.8	0%	0%	10173.2
115	+yg	-17114.6	0%	99%	10426.2
ME + gr + ob + rb + yb + yo + ro + co + go + yr + gb + cb + yg					
116	+cy	-17137.5	0%	86%	10877.5
117	+cg	-17156.1	0%	0%	10623.7
118	+cr	-17139.3	0%	14%	10548.4
ME + gr + ob + rb + yb + yo + ro + co + go + yr + gb + cb + yg + cy					
119	+cg	-17178.6	0%	0%	11074.2
120	+cr	-17162.2	0%	100%	10999.8
ME + gr + ob + rb + yb + yo + ro + co + go + yr + gb + cb + yg + cy + cr					
121	+cg	-17203.3	0%	100%	11196.3

Table E.2: Mean and standard deviation of coefficients for high probability accident severity models with factor year replaced by economy

		Model 11		Model 16		Model 30	
		Mean	SD	Mean	SD	Mean	SD
Constant		-2.69	0.17	-2.60	0.14	-2.72	0.17
Vehicle	4x4	-	-	-	-	-	-
	Large	0.35	0.09	0.36	0.09	0.36	0.09
	Medium	0.22	0.07	0.22	0.07	0.22	0.07
	Minis	-0.12	0.07	-0.11	0.07	-0.11	0.07
	Sports	0.34	0.09	0.35	0.09	0.35	0.09
	Small	0.04	0.07	0.04	0.07	0.04	0.07
Economy		0.03	0.01	0.03	0.01	0.03	0.01
Car age	0–2	-	-	-	-	-	-
	3-5	0.02	0.15	-0.01	0.05	0.01	0.15
	6-10	0.24	0.14	0.00	0.04	0.23	0.14
	11-15	0.17	0.16	0.01	0.05	0.16	0.16
	16+	0.37	0.23	0.08	0.07	0.38	0.23
Road	M	-	-	-	-	-	-
	A	-0.01	0.12	-0.11	0.05	-0.01	0.12
	Minor	-0.09	0.12	-0.27	0.06	-0.09	0.12
Overturn		0.13	0.03	0.22	0.04	0.22	0.04
Bend		0.21	0.03	0.28	0.04	0.28	0.04
Age &	0–2 & M	-	-			-	-
road	0–2 & A	-	-			-	-
	0–2 & Min	-	-			-	-
	3–5 & M	-	-			-	-
	3–5 & A	-0.03	0.17			-0.03	0.17
	3–5 & Minor	-0.03	0.17			-0.02	0.17
	6–10 & M	-	-			-	-
	6–10 & A	-0.23	0.15			-0.23	0.15
	6–10 & Minor	-0.31	0.15			-0.30	0.15
	11–15 & M	-	-			-	-
	11–15 & A	-0.07	0.17			-0.07	0.17
	11–15 & Minor	-0.29	0.18			-0.28	0.17
	16+ & M	-	-			-	-
	16+ & A	-0.36	0.25			-0.37	0.26
	16+ & Minor	-0.31	0.25			-0.32	0.25
Overturn	&Bend			-0.19	0.06	-0.19	0.06
Deviance		9109.05	6.74	9106.05	5.63	9099.72	6.90

Bibliography

af Wahlburg, A. and Dorn, L. (2007), "Culpable versus non-culpable traffic accidents; what is wrong with this picture?" *Journal of Safety Research*, 38, 453–459.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003), "An Introduction to MCMC for Machine Learning," *Machine Learning*, 50, 5–43.

Bolstad, W. M. (2011), *Understanding Computational Bayesian Statistics*, John Wiley and Sons.

Broughton, J. (2003), "The benefits of improved car secondary safety," *Accident Analysis and Prevention*, 35, 527–535.

— (2007), "Casualty rates by type of car," Tech. Rep. TRL Published Project Report PPR 203, Wokingham.

— (2009), personal communication.

Broughton, J. and Buckle, G. (2007), "Monitoring progress towards the 2010 casualty reduction target - 2006 data," Tech. Rep. TRL Published Project Report PPR668, Wokingham.

Broughton, J. and Knowles, J. (2010), "Monitoring progress towards the 2010 casualty reduction target - 2008 data," Tech. Rep. TRL Report TRL673, Wokingham.

Carr, B. (1970), "A statistical analysis of rural Ontario traffic accidents using induced exposure data," in *Proceedings. of the Symposium on the Use of Statistical Methods in the Analysis of Road Accidents*, OECD, France, pp. 86–92.

Chandraratna, S. and Stamatiadis, N. (2009), "Quasi-induced exposure method: evaluation of not-at-fault driver assumption," *Accident Analysis and Prevention*, 41, 308–313.

Chipman, H., George, E., and McCulloch, R. (2001), "The practical implementation of Bayesian Model selection," *IMS lecture notes - Monograph series*, 38, 65–134.

Cuerden, R., Pittman, M., Dodson, E., and Hill, J. (2008), "The UK On The Spot Accident Data Collection Study Phase II Report," Tech. Rep. Research Report No. 73, Department for Transport Road Safety, London.

Cullen, A. C. and Frey, H. C. (1999), *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*, Plenum Press, New York.

Deming, W. and Stephan, F. (1940), "On least square adjustment of sampled frequency tables when the expected marginal totals are known," *Annals of Mathematical Statistics*, 6, 427–444.

Department for Transport (2010), "How National Traffic Estimates Are Made," `http://www.dft.gov.uk/matrix/Estimates.aspx`.

— (2011), "Reported Road Casualties Great Britain: 2010 Annual Report," Tech. rep., The Stationary Office, London.

— (2012), "National Travel Survey 2002-2010," `http://www.esds.ac.uk/findingData/snDescription.asp?sn=5340`.

European Commission (2012), "Speed and Injury Severity," `http://ec.europa.eu/transport/road_safety/specialist/knowledge/speed`.

Fridstrom, L. and Ingebrigtsen, S. (1991), "An aggregate accident model based on pooled, regional time-series data," *Accident Analysis and Prevention*, 23, 363–378.

Gelfand, A. and Dey, D. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society, Series B*, 56, 501–514.

Gelfand, A. and Smith, A. (1990), "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Florida, USA: Chapman and Hall.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, 6, 733–807.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–511.

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

Geweke, J. (1992), in *Bayesian Statistics*, eds. J. M. Bernardo, J. O. Berger, A. P. D. and Smith, A. F. M., Oxford, UK: Clarendon Press, vol. 4, chap. Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments.

Gill, J. (2002), *Bayesian Methods for the social and behavioural sciences*, Florida, USA: Chapman and Hall.

Good, I. (1985), "Weight of evidence: a brief survey," in *Bayesian Statsitics 2*, eds. Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., Elsevier Science Publishers, pp. 249–270.

Green, M. J., Medley, G. F., and Browne, W. J. (2009), "Use of posterior predictive assessments to evaluate model fit in multilevel logistic regression," *Veterinary Research*, 40, 30–39.

Green, P. (1995), "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hammersley, J. M. and Clifford, P. (1971), "Markov fields on finite graphs and lattices,"
`http://www.measuringworth.org/datasets/ukgdp/result.php`.

Hart, A., Smith, G. C., Macarthur, R., and Rose, M. (2003), "Application of uncertainty analysis in assessing dietary exposure," *Toxicology Letters*, 140–141, 437–442.

Hastings, W. (1970), "Monte Carlo Sampling Methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Helton, J. (1996), "Probability, conditional probability and complementary cumulative distribution functions in performance assessment for radioactive waste disposal." Tech. Rep. SAND95-2571, UC-721, Sandia National Laboratories, Albuquerque.

Hotelling, H. (1931), "The generalization of Student's ratio," *Annals of Mathematical Statistics*, 2, 360–378.

Jeffreys, H. (1961), *Theory of Probability*, Clarendon Press, Oxford, 3rd ed.

Jiang, X. and Lyles, R. W. (2007), "Difficulties with quasi-induced exposure when speed varies systematically by vehicle type," *Accident Analysis and Prevention*, 39, 649–656.

Kass, R. and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Keall, M. and Newstead, S. (2007), "Four-wheel drive vehicle crash involvement risk, rollover risk and injury rate in comparison to other passenger vehicles: Estimates based on Australian and New Zealand crash data and on New Zealand motor vehicle register data," Tech. Rep. Monash University Accident Research Centre, Report No. 262, Melbourne, Australia.

Kelly, D. L. and Smith, C. L. (2009), "Bayesian inference in probabilistic risk assessment - The current state of the art," *Reliability Engineering and System Safety*, 94, 628–643.

Knowles, J., Broughton, J., and Buckle, G. (2007), "Sports Utility Vehicles: Collision Risks and Outcomes for Londons Road Users," Tech. Rep. TRL Published Project Report PPR 226, Wokingham.

Lardelli-Claret, P., Jiminez-Moleon, J. J., Luna-del Castillo, J. d. D., Garcia-Martin, M., Moreno-Abril, O., and Bueno-Cavanillas, A. (2005), "Comparison between two quasi-induced exposure methods for studying risk factors for road crashes," *American Journal of Epidemiology*, 163, 188–195.

Lloyd, L. and Forster, J. (in press), "Modelling Trends in Road Accident Frequency – Bayesian Inference for Rates with Uncertain Exposure," *Journal of Computational Statistics and Data*.

Lloyd, L. K., Reeves, C., Scoons, J., and Broughton, J. (2013), "Investigating the reduction in fatal accidents in Great Britain from 2007–2010," Tech. Rep. TRL Published Report PPR663, Wokingham.

Loomis, D. and Kromhout, H. (2004), "Exposure variability: concepts and applications in occupational epidemiology," *American Journal of Industrial Medicine*, 45, 113–122.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS – a

Bayesian modelling framework: concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

Lyles, R. (1994), "Quasi-induced exposure: to use or not to use?" in *73rd Transportation Research Board Annual meeting*.

Lyles, R., Stamatiadis, P., and Lighthizer, D. (1991), "Quasi-induced exposure revisited," *Accident Analysis and Prevention*, 23, 275–285.

Lynch, S. M. and Western, B. (2004), "Bayesian Posterior Predictive Checks for Copmlex Models," *Sociological Methods and Research*, 32, 301–335.

Mahalanobis, P. (1936), "On the generalised distance in statistics," in *Proceedings of the National Institute of Sciences of India*, vol. 2 (1), pp. 49–55.

Mardia, K. (1985), "Mardia's Test of Multinormality," in *Encyclopedia of Statistical Sciences*, eds. Kotz, S. and Johnson, N., Wiley, pp. 217–221.

Martz, H. F. and Picard, R. P. (1995), "Uncertainty in Poisson event counts and exposure time in rate estimation," *Reliability Engineering and Systems Safety*, 48, 181–190.

Meng, X.-L. and Wong, W. H. (1996), "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration," *Statistica Sinica*, 6, 831–860.

Mengersen, K. and Tweedie, R. (1996), "Rates of convergence of the Hastings and Metropolis algorithms," *Annals of Statistics*, 24, 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., and Teller, A. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

Mitra, S. and Washington, S. (2007), "On the Nature of Over-dispersion in Motor Vehicle Crash Prediction Models," *Accident Analysis and Prevention*, 39, 459–468.

Molitor, J., Molitor, N.-T., Jerrett, M., McConnell, R., Gauderman, J., Berhane, K., and Thomas, D. (2006), "Bayesian modeling of air pollution health effects with missing exposure data," *American Journal of Epidemiology*, 164, 69–76.

Morgan, M. G. and Henrion, M. (1990), *Uncertainty: A guide to dealing with uncertainty in quantitativerisk and policy analysis*, Cambridge University Press, New York.

Officer, L. H. and Williamson, S. H. (2010), "What was the UK GDP then?" `http://www.measuringworth.com/ukgdp` .

Preller, L., Kromhout, H., Heederik, D., and Tielen, M. (1995), "Modeling long-term average exposure in occupational exposure-response analysis," *Scandinavian Journal of Work and Environmental Health*, 21, 504–512.

Qin, X., Ivan, J. N., Ravishanker, N., Liu, J., and Tepas, D. (2006), "Bayesian estimation of hourly exposure functions by crash type and time of day," *Accident Analysis and Prevention*, 38, 1071–1080.

Queen, C. and Albers, C. (2009), "Intervention and Causality: Forecasting traffic flows using a dynamic Bayesian network," *Journal of the American Statistical Association*, 104, 669–681.

Redondo-Calderon, J. L., Luna-del Castillo, J. d. D., Jimenez-Moleon, J. J., Garcia-Martin, M., Lardelli-Claret, P., and Galvez-Vargas, R. (2001), "Application of the induced exposure method to compare risks of traffic crashes among different types of drivers under different environmental conditions," *American Journal of Epidemiology*, 153, 882–891.

Robert, C. and Casella, G. (2010), "A History of Markov Chain Monte Carlo - Subjective Recollections from Incomplete Data," in *Handbook of Markov Chain Monte Carlo: Methods and Applications*, eds. Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., Florida, USA: Chapman and Hall.

Roberts, G., Gelman, A., and Gilks, W. (1997), "Weak convergence and optimal scaling of random walk Metropolis algorithms," *The Annals of Applied Probability*, 7, 110–120.

Rosas-Jaimes, O. A., Campero-Carmona, A. C., and Snchez-Flores, O. L. (2011), "Prediction under Bayesian approach of car accidents in urban intersections," in *3rd International Conference on Road Safety and Simulation*.

Seixas, N. S. and Sheppard, L. (1996), "Maximizing accuracy and precision using individual and grouped exposure assessments," *Scandinavian Journal of Work and Environmental Health*, 22, 94–101.

Shapiro, S. S. and Wilk, M. B. (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591–611.

Smith, A. and Roberts, G. (1991), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society: Series B*, 55, 3–23.

Sohn, M. D., McKone, T. E., and Blancato, J. N. (2004), "Reconstructing population exposures from dose biomarkers: inhalation of trichloroethylene (TCE) as a case study," *Journal of Exposure Analysis and Environmental Epidemiology*, 14, 204–213.

Sonkin, B., Edwards, P., Roberts, I., and Green, J. (2006), "Walking, cycling and transport safety: an analysis of child road deaths," *Journal of the Royal Society of Medicine*, 99, 402–405.

Stamatiadis, N. and Deacon, J. A. (1996), "Quasi-induced exposure: methodology and insight," *Accident Analysis and Prevention*, 296, 37–52.

Starnes, M. and Longthorne, A. (2003), "Child pedestrian fatality rates by striking vehicle body type: A comparision of passenger cars, sports utility vehicles, pickups and vans." Tech. Rep. Traffic Safety Facts research Note, NHTSA NCSA.

Tanner, M. and Wong, W. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–550.

Thorpe, J. (1967), "Calculating relative involvement rates in accidents without determining exposure," *Traffic Safety Research Review*, 11, 3–8.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1786.

Tierney, L. and Kadane, J. (1986), "Accurate Approximation for Posterior moments and marginal densities," *Journal of the American Statistical Association*, 81, 82–86.

Tunaru, R. and Jarrett, D. (1998), "Analysis of causality of road accident data using graphical models," in *Proceedings of the third IMA conference Maths in Transport Planning and Control*.

Van Belle, G. (2008), *Statistical Rules of Thumb*, Wiley-Blackwell.

Van den Bossche, F. and Wets, G. (2003), "A Structural Road Accident Model for Belgium," Tech. Rep. RA-2003-21, Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit.

Yannis, G., Golias, J., and Papadimitriou, E. (2005), "Driver age and vehicle engine size effects on fault and severity in young motorcyclists accidents," *Accident Analysis and Prevention*, 37, 327–333.

— (2007), "Accident risk of foreign drivers in various road environments," *Journal of Safety Research*, 38, 453–459.