MODELLING ROAD ACCIDENTS FROM NATIONAL DATASETS: A CASE STUDY OF GREAT BRITAIN

Abdul Qadeer Memon

A thesis submitted to University College London for the degree of Doctor of Philosophy

Centre for Transport Studies
University College London

July 2012

DECLARATION

I, *Abdul Qadeer Memon,* confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abdul Qadeer Memon

ABSTRACT

This study investigates the occurrence of road traffic accidents in Great Britain at a national scale. STATS 19 data for road accidents, vehicles involved in road accidents and casualties occurring over several years were analysed and modelled using various statistical techniques. The main aims of this research were to investigate the use of different statistical model formulations and to investigate the numbers of road accidents, casualties, and vehicles involved that occur on each day. Generalized linear model (GLM), generalized estimation equation (GEE), and hierarchical generalized linear model (HGLM) formulations were investigated for this purpose. The variables of weekday 3 (weekday, Saturday, Sunday), seasons (Spring, Summer, Autumn, Winter), month, time, Public holidays, Christmas holidays, new-year holidays, road type and vehicle class, together with certain interactions between them, were found to be important in developing models of risk per unit of distance travel. Additional variables of distance travelled per vehicle, vehicles per head of population, population density, meteorological factors were also investigated, and population, age group and gender were used to develop models of casualty rate per person-year.

The GLM model structure with log link function was found to fit data for the occurrence of road accidents reasonably well when the negative binomial distribution was adopted to accommodate over-dispersion beyond Poisson levels. The GEE with negative binomial error together with autoregressive (AR1) structure was preferred over the GLM as it can also accommodate serial correlation that was found to be present in the data due to the natural order of the observations. The coefficients and significance levels of some variables were found to change significantly if the presence of serial correlation is not respected. Finally HGLM with Poisson-gamma errors and log link function was used to estimate the number of casualties involved in road accidents on each day. The advantage of HGLM over GLM and GEE is that it can account for variability within and between clusters using both random effects and dispersion modelling: this was found to be substantial. However, unlike GEE, HGLM cannot accommodate time series structure so that the coefficients and the associated standard errors of some of the variables should be viewed with caution.

From the model results, it is found that distance travelled provided a good measure of exposure to risk in most cases, and that each of distance travelled per vehicle, population density and rain is associated with greater risk for road accident per unit of travel whereas

risk diminishes with increase in each of numbers of vehicles per person and mean minimum monthly temperature. The risk per unit of travel was also estimated for each of 5 classes of vehicles on each of 5 different kinds of roads. Finally the age and gender specific rate of casualty per person-year was estimated for each combination of age group and gender. The results obtained from this study will lead to the promotion of safe usage of road and vehicle class combinations by raising travellers' awareness. On the other hand the casualty rates estimated for each of the 8 age groups and two gender groups by vehicle class will help to identify those that need more attention. These results will help various educational, planning, and rescue agencies to identify target groups for education and engineering initiatives to improve road safety.

ACKNOWLEDGEMENTS

First of all, the author would like to express deepest sense of gratitude to supervisor Professor Benjamin Heydecker for his precious guidance, constant encouragement and excellent advice throughout this study. The author will remain forever indebted for his efforts that he had put not only in getting this study but also for enhancing the learning skills of author. The author is also immensely grateful to the secondary supervisor Dr Helena Titheridge for her guidance, useful discussions, and valuable suggestions on the subject. Grateful acknowledgments are extended to Professor Roger Payne, Professor Youngjo Lee, and John Shrewsbury for their valuable suggestions and comments on HGLM and GenStat software.

The author is also grateful to his fellow students at Centre for Transport Studies, UCL for providing excellent working environment.

Grateful acknowledgments are extended to Commonwealth Scholarship Commission, United Kingdom for providing financial support for PhD at Centre for Transport Studies, University College London.

The author understands and acknowledges the contribution made by his grandparents late Mr Rasool Bux and late Mrs Rasool Bux for their untiring efforts and dedication to give best quality education to their children. The author also pays high regards to all his family members especially father late Prof. Dr. Abdul Ghani Memon, mother Shamim Akhtar, sister Asifa Memon, sister Faiza Memon and Wife Beenish Fatima Memon for their good wishes and constant encouragement. The author also want to appreciate the sacrifices made by his children Faris Ahmed Memon and Azlan Ahmed Memon who spent most of their early days waiting for the author while he was working hard for the completion of this study.

The author dedicates this thesis to his father Late Professor Dr Abdul Ghani Memon. Without his support, guidance, encouragement, help, love and prayers this would not have been possible. Although Baba is not here today to share this great news of achieving this milestone and to see one of his dream fulfilled. By dedicating this achievement to BABA I just want to say that our dreams have been fulfilled BABA and I LOVE YOU VERY MUCH.

Contents

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

In the light of the particular usage in this thesis of certain terms, the following glossary is provided to clarify this.

Circumstantial variables: These variables represent the characteristics of transport activity in a region in a sacle-free way. The variables of population density, number of vehicles per head of population, number of vehicles per kilometre of road length, number of vehicles per square kilometre of surface area and ratio of each road class to total road length are termed as circumstantial variables.

Risk: measure of accident involvement per vehicle kilometre of distance travelled.

Rate: measure of accident involvement per person-year.

1. INTRODUCTION

1.1 GENERAL BACKGROUND

Every year more than a million people die in road traffic accidents worldwide, and 50 million are injured. This is likely to increase by 65 percent over the next 20 years due to rapid increase in motor vehicle ownership and usage in large developing countries. For this reason, traffic accidents are one of the world's largest public health problems. The problem is all the more acute because the victims are overwhelmingly young and healthy prior to accidents (World Health Organization, 2004). According to World Health Organization (WHO) projections, by 2020 road traffic accidents will account for 2.3 million deaths worldwide, with over 90 percent occurring in low and middle income countries.

Road safety is one of the main issues in transportation. In many higher income countries the number of road fatalities has decreased in the last 20-25 years due to the application of systematic approach to improve road safety (International Traffic Safety Data and Analysis Group, 2008). The Organisation for Economic Co-operation and Development (OECD) countries, which include most of the industrialised countries, have achieved considerable success in improving road safety by applying proper road accident countermeasures including education, engineering and enforcement. In industrialised countries, availability of accurate road accident data is regarded as an essential starting point for this work. By using available road accident data, suitable remedial measures can be devised and appropriate strategies planned by identifying the key target groups for reducing road accidents. The data of the 29 member countries of OECD, which is available from the International Traffic Safety Data and Analysis Group in the form of the International Road Traffic and Accident Database (IRTAD), show reduction of about 12 percent in road fatalities in 2008 by comparison to 2005. The latest data released by IRTAD (2010) shows that in 2008 Spain, Israel, Denmark, United Kingdom and Slovenia achieved substantial reductions in the number of road fatalities.

In Great Britain substantial reduction from 5,953 road accident deaths in 1980 to 1,850 in 2010 is observed (Department for Transport, 2011). Successive UK Governments have committed substantial efforts and resources to reduce the number of road accidents and casualties by increasing awareness among people and by applying safety intervention

programmes across the whole country. According to the 2008 OECD data, Great Britain is considered to have a good road safety record as it is ranked 3rd in the OECD countries for having the lowest number of persons killed per million population in road accidents. There were only 4.3 persons killed per 100,000 population and 5 persons killed per billion vehicle kilometres of travel. Iceland and Netherlands were found to be safer per head of population whilst Iceland has the lowest number of road accident deaths per billion vehicle kilometres. The comparison of the deaths per 100,000 population of the OECD countries is shown in Figure 1.1, which shows that scope still exists for further effort to reduce the number of road accidents in Great Britain.

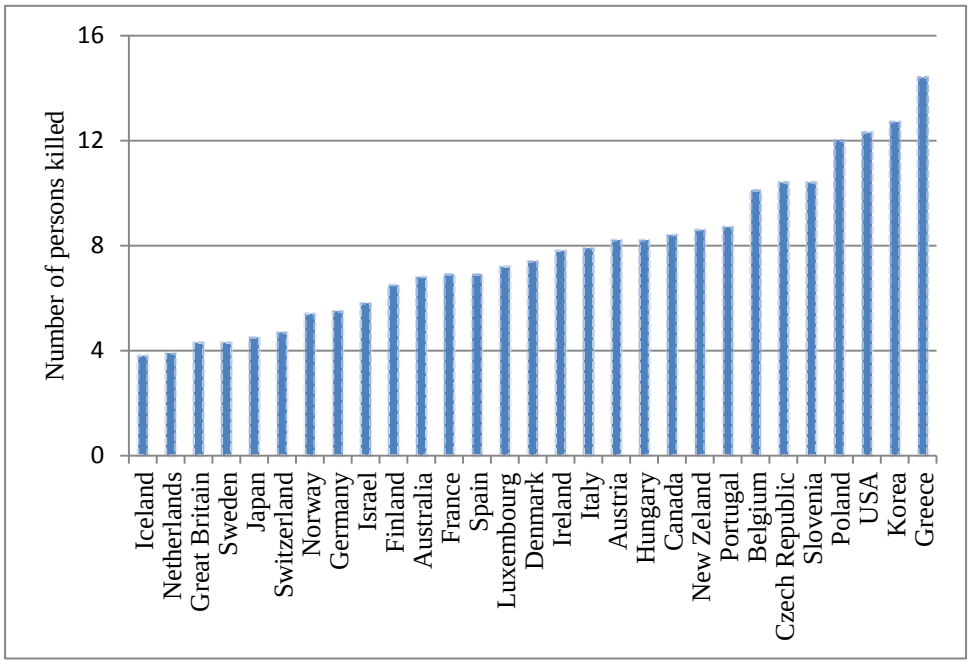
1.2 NEED TO STUDY ROAD SAFETY

Road safety research is the scientific study of road and traffic systems with the main aim of finding ways of reducing the number of road accidents and their severity. It is also of considerable importance to the economy of the country. In economic terms the cost of road traffic injuries is estimated to be 1 percent of the gross national product (GNP) of low income countries, 1.5 percent in middle income countries and 2 percent in high income countries. It was also estimated that the global cost of road traffic accidents is $518 billion per year (Jacobs, 2000). In Great Britain 1,730 fatal accidents, 20,440 serious accidents, and 132,243 slight accidents were reported in 2010. The total benefit value of prevention of personal injury road traffic accidents was estimated to be £10.6 billion. In addition to this, there were 2.3 million damage-only accidents valued at a further £4.4 billion. Hence the total value of the prevention of all road accidents in 2010 was estimated to be £14.9 billion based on 2009 prices and values (Department for Transport, 2011).

1.3 MATHEMATICAL MODELLING OF ROAD ACCIDENTS

The number of road accidents can be modelled by using various techniques to identify the relationship of different variables with number of road accidents so that insights can be obtained for improving road safety and suitable safety intervention programmes can be developed. This section gives an overview of the techniques that have been used by various researchers for modelling the number of road accidents and problems this entails.

Figure 1.1: Number of people killed per 100,000 population in OECD countries (2008)



Source of data: International road traffic and accident dataset (2010)

1.3.1 Multiple regression, Poisson, and negative binomial regression

In earlier research, relationships between road accidents and other variables have been estimated by using the conventional ordinary least square multiple regression techniques. This method assumes that the dependent variable is continuously and normally-distributed with a constant variance. The conventional multiple linear regression technique lacks the distributional property necessary to adequately describe random, discrete, and non-negative events such as road traffic accidents. Various authors including Miaou (1993), and Miaou and Lum (1993) have shown that the test statistics derived from these models are not always reliable. In other studies by Maycock and Hall (1984), Hall (1986), Hadi et al (1995), and Anis (1996) significant advances have been made to describe traffic accident count data and to produce more accurate and reliable models through the use of Generalized Linear Models (GLMs) with log-linear form, and Poisson and negative binomial distributions.

Maher and Summersgill (1996) found that variance of the count data is generally higher than the mean. The extra variation is known as over-dispersion. When using Poisson regression in the presence of over-dispersion, model parameter estimates will still be close to their true values but their variance of estimation will tend to be underestimated and the significance

levels of the estimated coefficients will therefore be overstated. In order to overcome the over-dispersion problem Abdel-Aty and Radwan (2000), Guevara et al (2004), and McCarthy (2005) among many others have adopted the negative binomial distribution which allows the variance to exceed the mean.

1.3.2 Problems with count/ panel/ national accident datasets

According to Sittikariya and Shankar (2005) two important issues that arise in the analysis of count data of this kind are serial correlation, which arises because the data are in time series, and excessive zeros. Time-series and repeated observations of multiple years of cross-sectional data on road accident occurrence are often available in the public domain, including time-series information on traffic volumes, road accident counts, and roadway geometrics. This then conforms to repeated observations of several random variables and hence to the concept of panel data.

In modelling the frequency of road traffic accidents, both of these two problems may occur. Researchers have adopted various techniques to address them.

1. In the presence of repeated observation effects or serial correlation, the efficiency of parameter estimates comes into question. Wang and Abdel-Atey (2006), and Lord and Persaud (2000) used Generalized Estimating Equations (GEE) to accommodate serial correlation in data for modelling the number of road accidents.

2. The presence of excess zeros in the data may also lead to inaccurate results. This problem was solved by zero inflated Poisson and zero inflated negative binomial models by Shankar et al (1997). This technique deals with over-dispersion that can arise due to excessive zeros from many sites at which no accidents are observed.

1.4 DATA REQUIRED TO STUDY ROAD SAFETY

The availability of accurate and comprehensive data related to road accidents can promote improvements in road safety. The interpretation of the data can lead to better identification and understanding of problems, and hence will assist in developing and evaluating appropriate road safety remedial measures. Road safety professionals require information

about large numbers of road accidents to identify hazardous locations or to identify groups of people who are at higher risk of being involved in road accidents. This will lead to the formulation of plans to improve road safety for target locations and groups.

The need and importance of having road accident data prompts authorities to design road accident data collection, management, and retrieval system for road accident data. Transport authorities are responsible in most countries to decide which types of data to be collected, coded and managed in the database. The following information is typical of that collected by authorities to describe road traffic accidents:

- ***Where*** the road accident occurred: road name, road classification, type of traffic control, location coordinates;
- ***When*** the road accident occurred: Time of day, day of the week, month, year;
- ***Who*** was involved: vehicles, people, roadside objects;
- ***What*** was the result of the road accident: fatal, personal injury, property damage;
- ***How*** the road accident occurred.

The road accident data can be used by many professionals in various ways. In general, potential users of road accident data will include the following:

- Road safety engineers for the purpose of improving elements of the road network and developing remedial traffic measures;
- Groups that have responsibility of improving road safety education;
- Police in relation to enforcement activities such as the location of officer patrols and speed cameras and other priorities;
- Researchers may need to conduct rigorous investigation to identify target locations, activities, and groups;
- Lawyers for compensation for injuries and other losses;
- Vehicle and infrastructure manufacturers may wish to assess the safety performance of their product.

The most widely available source of road accident data is based upon police report forms. In most countries the site of the road accident is attended by a police officer, which results in the

production of a road accident report. Road accidents do not always fit standard formats so that a road accident report form will not always describe completely every road accident that has occurred. The training and motivation as well as experience and skills of the police officer are also important in recording the details accurately. Notwithstanding this, police reports remain the best source of national road accident data in most countries. Data obtained from police reports generally inform us about the where, when, who and what questions but tells us little about how and why the road accident occurred. In some countries such as Great Britain there is also some additional information available that can lead to an understanding of the contributory factors involved in a road accident: since 2005, a choice of up to 6 factors from a range of 76 have been recorded for each road accident as part of the British STATS 19 national data system for road accidents reported at scene by the police. Each factor is associated with one of nine groups that are mostly classified according to the three elements: road environment, vehicle defects, and user (Department for Transport, 2011).

These road accident recording datasets are available in various countries but their potential use in modelling road accidents at national level has rarely been explored. The road accident models developed by using these datasets will help to summarise national trends in road accident occurrence. National and local authorities can use these models to identify important factors that contribute to road accidents and appropriate target groups for attention. Remedial policies can then be developed accordingly. The development of road accident prediction models from national road accident datasets can lead to better understanding of the road accidents. This research opportunity is developed in the present thesis with the ultimate intention to help improve road safety policy and practice in Great Britain and with the possibility of transferring the resulting methodology to other countries.

1.4.1 Road accident reporting system in Great Britain

In Great Britain, police complete a STATS 19 form (Department for Transport, 2010) for each road accident involving personal injury that occurred on a public highway and that becomes known to the police within 30 days of its occurrence. Personal injury road accidents statistics were first collected in 1909, and the new system of collecting information known as STATS 19, was introduced in 1949. Information about the road accident, vehicles involved, and casualties is collected. Data is collected each month from police forces throughout the year. Road accidents are coded by local authorities and sent to the Department for Transport

which compiles and maintains data. The results are published for local authorities, police forces, regions, and for Great Britain. These results are used extensively for research to influence road safety improvements. STATS 19 data is also extensively used by the following organisations:

1. Department for Transport (DfT), Scottish Executive (SE) and National Assembly for Wales (NAfW) annual statistics on road accidents and casualties;
2. In local authorities engineers use STATS 19 data to identify priority sites for remedial measures;
3. Road safety officers develop national and local education and training programmes based on evidence gathered from the data; and
4. The police use these data for tactical deployment of patrols in order to reduce the number of casualties.

1.4.2 United Kingdom road safety plans

The UK Department for Transport (DfT) has the responsibility for developing road safety policy of the United Kingdom. The UK road safety strategy is comprehensive, covering ten priority themes which are: safer for children, safer drivers (training and testing), safer drivers (alcohol, drugs and drowsiness), introduce new measures to reduce drink-driving, develop more effective ways to tackle drug-driving, safer infrastructure, safer speeds, safer vehicles, better enforcement and promoting safer road use (Department for Transport, 2010).

By 2010, the UK government planed to achieve, compared with the average for 1994-98;

- 40 percent reduction in number of people killed or seriously injured in road accidents;
- 50 percent reduction in numbers of children killed or seriously injured; and
- 10 percent reduction in slight casualty rate, expressed as the number of people slightly injured per 100 million vehicle kilometres;

By comparing the 2010 road accident data (DfT, 2011) with the 1994-98 average, it is observed that in 2010:

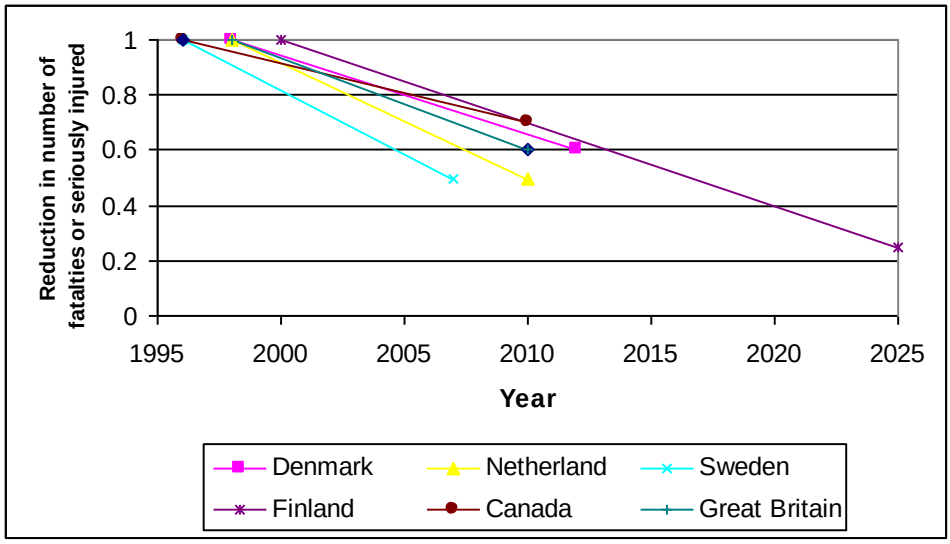
- The number of persons killed was 48 percent lower;
- The number of children killed or seriously injured was 64 percent lower;
- The slight casualty rate was 32 percent lower;
- In contrast the traffic rose by 13 percent over this period.

From this, we see that the 2010 annual data met all of the casualty reduction targets that were set for year 2010.

The Department for Transport also evaluates the road safety programme. Routine monitoring is carried out annually, and formal programme reviews are planned to be carried out every three years. General monitoring indicators are: the number of road accidents and casualties by severity and by road user group, drink-driving, use of seatbelts, use of cycle helmets, speed, road user attitudes by means of surveys, and other ad hoc surveys. Other indicators that are monitored are: traffic volume by vehicle type, travel patterns, modal split, vehicle registrations, driving test volumes and pass rates. Cost/benefit studies of various measures are an integral part of programme evaluation.

The road accident data systems and road safety improvement plans of several OECD countries are described in Appendix A1.1 and A1.2. Figure 1.2 shows the comparison of the road safety targets of some of the OECD countries which indicate that most of the countries had targets of a 40 percent or higher reduction in the number of fatal and serious injuries. From Figure 1.2, it can be seen that Great Britain had target of a 40 percent reduction in the number of fatal or serious injuries by 2010 from the base year average of 1994-1998 whereas Finland had a target of a 75 percent reduction in fatalities by 2025 from the base year of 1996.

Figure 1.2: Comparison of the road safety targets of some of OECD countries



Source of data: International road traffic and accident dataset (2010)

1.5 AIMS AND OBJECTIVES OF THE RESEARCH

The aim of the present research is to develop road accident prediction models that can describe and accurately estimate the number of road accidents, casualties, and vehicles involved in road accidents in Great Britain at an aggregate (national) or at a disaggregated level, such as police force area, by using the national road accident dataset of STATS 19. Statistical models of this kind embody the relationship between number of road accidents and other variables such as day of week, month, time, holidays, total distance travelled, number of vehicles per head of population, population density, road class, vehicle type, age, gender, and various meteorological factors. These relationships can thereby be explored and better understood. It is to be noted that the selection of these variables is limited due to the nature of the STATS 19 data, although information from other datasets (national travel survey data, population and meteorological data) that are available at national scale are also used here.

A methodological aspect of this research is to identify suitable techniques to model the number of road accidents occurring on each day by combining the data available in the accident, casualty, and vehicle sections of STATS 19 along with other related available data. From the results, the risk per unit of exposure can also be estimated which can be used to identify target groups for improvements in road safety. By examining the safety record of different kinds of vehicle on different kinds of road and the corresponding amount of use, the range of risks of different road and mode usage combinations can be estimated. The results of

this research will support advice to travellers and will help various planning and rescue agencies to develop road safety intervention programmes. This will also enable agencies to allocate their resources in a better way by anticipating how many road accidents are likely to occur on each day throughout the study area for various road classes, vehicle classes, gender, and age groups.

The following specific objectives are identified for the research of this thesis:

1. To investigate the number of road accidents on each day in Great Britain, the casualties incurred and vehicles involved. This will involve combined use of the national road accident dataset (STATS 19), national travel survey data (NTS), population data, and meteorological data of Great Britain.

2. To determine the relationship between number of road accidents and different variables available in accident, casualty, and vehicle sections of the STATS 19 dataset.

3. To evaluate the performance of various statistical models developed according to the principles of the generalized linear model (GLM), generalized estimation equation (GEE) and hierarchical generalized linear model (HGLM), and based on this to identify the properties and relative merits of these modelling approaches.

4. To compare the risk per unit of distance travelled for different combinations of vehicle class and road type, and casualty rate per person-year for gender and age group.

1.6 STRUCTURE OF THE THESIS

In this thesis a range of statistical modelling techniques are considered that are used to estimate the numbers of road accidents, vehicles involved and casualties. This entails analysis of different outcomes, reported by different response variables giving the number of road accidents, vehicles and casualties occurring during various time periods such as day of the week or month. In chapters 2 and 3, the number of road accidents is used as the response variable whilst data from the national travel survey (NTS), population data, and meteorological data are used jointly as explanatory variables. Important information about the

road type, vehicles class, age and gender of casualties is included in STATS 19 data, but not within the accident section. In order to use this information, vehicles and casualty sections of STATS 19 data are combined with information in the accident section. Due to this, the number of vehicles involved in road accidents on each day is used as response variable in Chapter 4 and the number of casualties in road accidents on each day in Chapter 5.

The modelling techniques used for this start from Generalized Linear Model (GLM) with Poisson and negative binomial regression, which is well established. After this, more advanced modelling methods are investigated. These are Generalized Estimation Equation (GEE) with auto-regressive (AR1) error structure to account for serial correlation, and the Hierarchical Generalized Linear Model (HGLM) with Poisson-gamma distribution which allows for the joint modelling of mean and dispersion. The purpose of this is to investigate the benefits of different methods and identify the scope and reliability of these models. In this thesis the datasets used are shown in Table 1.1 and the statistical modelling techniques are shown in Table 1.2.

This thesis is organized in six chapters. Tables and figures are presented in the body of the text where appropriate. Fully detailed results from the selected models are presented in appendices.

This first Chapter has introduced the background, aims and objectives, and provides an overview of the study.

Chapter 2 provides a background for modelling the number of road accidents occurring on each day in Great Britain at national and police force levels. The STATS 19 National accident dataset from 1991 to 2005 is used for this study. In addition to this, various other datasets such as population and population density obtained from the UK Statistics Authority, total distance travelled, number of vehicles, length of various road classes that were obtained from the Department for Transport (DfT) has been used. Variables derived from these were included into models to characterise transport activity of a region rather than describe its size. Two different datasets are prepared, each representing the road accidents on each day in Great Britain and in each of the 51 police force areas, each of which represents a group of local authorities. A Generalized Linear Model (GLM) with each of Poisson and negative binomial regression, and Generalized Estimation Equation (GEE) having auto-regressive

(AR1) negative binomial error structure is used to model these road accidents. Comparison of the results estimated by these techniques was carried out to explore which technique is appropriate for the data. The explanatory variables used for this are weekday 3 (weekday, Saturday, Sunday), season (Spring, Summer, Autumn, Winter), interaction of weekday 3 and season, month, time, Public holidays, Christmas holidays, new-year holidays, distance travelled per vehicle, population density and vehicles per head of population. The total distance travelled on each day is used as an offset variable to represent the exposure to risk.

Chapter 3 analyses the effect of meteorological factors on the occurrence of road accidents. The meteorological data obtained from Meteorological Office, UK, is used jointly with the STATS 19 accident data for the period 1991 to 2005. The numbers of road accidents occurring each month in 17 police force areas is used due to limitations on the availability of meteorological data. The Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) having AR1 error structure with negative binomial regressions are used for this. The explanatory variables used were month, time, population density, vehicles per head of population, mean minimum monthly temperature and amount of monthly rainfall. Total distance travelled in a month is used as offset to represent the exposure to risk.

Chapter 4 extends the use of the STATS 19 national accident dataset by linking the accident section with the vehicle information section for the years 2001 to 2005 to add road and vehicle type information. The numbers of vehicles involved in road accidents each day on each road class are extracted from the combined dataset produced for this study. Information about vehicle kilometres travelled by each group is obtained from the Department for Transport (DfT). The Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) having AR1 error structure with negative binomial regressions are used to estimate the number of vehicles involved in road accidents on each day. The variables of road class, vehicle class, weekday 4 (with 4 levels), season, interaction of weekday 4 and season, month, time, Public holidays, new-year holidays, Christmas holidays, interaction of road type and vehicle class, and variable representing the leisure motorcycling (MC-Rural-Sunday) are used. The distance travelled on each day was adopted for use as offset in these models to represent the exposure to risk.

Chapter 5 further extends the use of STATS 19 by joining the accident section with the casualty information section for years 2001 to 2005 to add age and gender information.

Casualties of all classes and severities were considered. This combination enables the addition of the parameters of gender, age group, and vehicle class to model the number of casualties in road accidents on each day across whole of Great Britain. The information about the population of each group was obtained from UK Statistics Authority whereas distance travelled by each age group by gender and vehicle class was obtained from the DfT. Five different datasets are used, each representing casualties by vehicle class. The Generalized Estimation Equation (GEE) having AR1 error structure and Hierarchical Generalized Linear Model (HGLM) methods are used to estimate the number of road casualties occurring each day by gender, age group and vehicle class. The variables of age group, gender, interaction of age group and gender, day of week, month, time, Public holidays, new-year holidays and Christmas holidays are used. In these models, population is used as an offset. The results estimated by GEE-AR1 and HGLM techniques are compared.

Chapter 6 summarises all the findings, draws some conclusions in respect of statistical methodology that has been used here and also in respect of road safety in Great Britain, discusses the implications for road safety research and policy. This leads to the identification of possibilities for future work.

Table 1.1: Datasets used in this Thesis

Dataset	Chapter	Description	No of observations	Time period
1	2	Number of road accidents on each day in Great Britain	5,479	1991-2005
2	2	Number of road accidents on each day in each of 51 police forces of Great Britain	279,429	1991-2005
3	3	Number of road accidents during each month in each of 17 police forces of Great Britain	3,060	1991-2005
4	4	Number of vehicles involved in road accidents on each day by road and vehicle combination	43,824	2001-2005
5	5	Number of (Car) casualties involved in road accidents on each day by age and gender combination	29,216	2001-2005
6	5	Number of (Walking) casualties involved in road accidents on each day by age and gender combination	29,216	2001-2005
7	5	Number of (Bicyclist) casualties involved in road accidents on each day by age and gender combination	29,216	2001-2005
8	5	Number of (Motorcyclists) casualties involved in road accidents on each day by age and gender combination	29,216	2001-2005
9	5	Number of (Bus) casualties involved in road accidents on each day by age and gender combination	29,216	2001-2005

Table 1.2: Models used in this Thesis

Chapter	Model	Description	Features
2,3,4	GLM	Log-Linear Poisson	
		Log-Linear negative binomial	Allows for over-dispersion
	GEE	GEE with auto regressive AR1	Accommodates serial correlation
5	GEE	GEE with auto regressive AR1	Accommodate serial correlation
	HGLM	Log-Linear Poisson and Gamma random effects	Includes random effects and models variance

GLM (Generalized Linear Model), GEE (Generalized Estimation Equation), HGLM (Hierarchical Generalized Linear Model)

2. MODELLING ROAD ACCIDENTS OCCURRENCE

2.1 INTRODUCTION

Road accidents are complex events involving the interaction of many factors (RoSPA, 2007). These factors include roads, vehicles, drivers, traffic, and environment. A lot of research has been done relating the number of road accidents to traffic flow and the geometric condition of the road. However, fewer attempts have been made to relate the number of road accidents to the day of week and month of year to find their effect on the occurrence of road accidents (Fridstrom et al, 1995; Leveine, et al, 1995).

In this chapter, we explore the relationship in the national data between road accidents, distance travelled, timing and circumstances of the road accidents as recorded in STATS 19 data. The Generalized Linear Model (GLM) (Nelder and Wedderburn, 1972) with each of Poisson and negative binomial regression and the Generalized Estimation Equation (GEE) (Liang and Zeger, 1986) having auto-regressive (AR1) error structure with negative binomial are used to model the number of road accidents occurring on each day in Great Britain and the results obtained by these are compared.

The Department for Transport, Local Government and Regions report (2001) and analysis of the road accidents data in STATS 19 format for Great Britain are shown in Table 2.1. This shows that there is no simple relationship between road accident frequencies and amount of travel as measured in either number of trips or distance travelled. It is seen that despite having highest number of road accidents per day, November does not have the highest exposure to risk as represented by either number of trips or distance travelled. In August fewer road accidents occur but greater distances are travelled, mainly for holiday and day-trip purposes, whereas school holidays at this time mean much less distance is travelled for education purposes (DTLR, 2001).

Table 2.1 further shows that weekdays have a greater number of road accidents than weekend days, and although they have a greater number of trips per day, they have less distance travelled than on weekend days. At weekends, greater distances are travelled for shopping and entertainment/public activity or day trip purposes (DTLR, 2001). From this information it can be seen that the number of road accidents occurring is not proportional to either the total

number of trips made or distance travelled during that time period. From this, we conclude that the risk of road accident occurrence varies according to circumstances whether it is measured by trip or by distance travelled.

Table 2.1: Trips made, distance travelled, and number of road accidents (1992-2000)

Average number of road accidents, average trips and average distance travelled on each day by month of the year, 1992-2000

	Jan	**Feb**	**Mar**	**April**	**May**	**June**	**July**	**August**	**Sept**	**Oct**	**Nov**	**Dec**
Average number of road accidents /day	605 (11)	608 (9)	599 (12)	606 (10)	625 (8)	646 (5)	641 (6)	628 (7)	664 (4)	685 (2)	729 (1)	670 (3)
Trips made /day	2.65 (11)	3.07 (1)	2.84 (8)	2.83 (9)	2.90 (= 5)	3.03 (2)	2.90 (= 5)	2.74 (10)	3.0 (3)	2.90 (= 5)	2.96 (4)	2.54 (12)
Average distance travelled / day (km)	24.16 (12)	26.65 (10)	27.2 6 (9)	30.87 (4)	29.87 (6)	30.6 (5)	31.58 (2)	33.35 (1)	31.17 (3)	29.29 (7)	27.9 (8)	25.84 (11)

Average number of road accidents, average trips and average distance travelled on each day of the week, 1992-2000

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
Average number of road accidents / day	638 (5)	650 (4)	658 (3)	679 (2)	755 (1)	625 (6)	489 (7)
Trips made / per day	2.86 (= 5)	3 (= 2)	3 (= 2)	3 (= 2)	3.14 (1)	2.86 (= 5)	2.14 (7)
Average distance travelled/ day (km)	27.43 (= 6)	27.43 (= 6)	28.29 (= 4)	28.29 (= 4)	31.57 (2)	32.43 (1)	28.57 (3)

Source of data: Department for Transport (2001)
**Figures in brackets show ranking*

The variation in number of road accidents occurring on different days of the week and month as shown in Table 2.1 and analysis of STATS 19 data emphasizes that detailed research is required to develop a model that can accurately describe the number of road accidents occurring on each day. With this approach, important variables affecting the number of road

accidents can be identified. This will help to establish how the number of road accidents in Great Britain can be reduced so that suitable safety intervention programmes can be developed accordingly by planning organisations. So an investigation of the occurrence of road traffic accidents at the national scale was carried out in the present study to:

- To identify the relationship between the number of road accidents and variables available in the national dataset;
- To identify those variables associated with the variation in number of road accidents; and
- To evaluate the performance of various statistical modelling formulations.

This chapter is organized as follows. Section 2.2 reviews the literature about the Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE). Section 2.3 briefly describes the data used for this study. Section 2.4 briefly analyses the data. Section 2.5 presents the process of model development and the basic structure of the model. Section 2.6 shows the model selection process, the results obtained from the developed models, goodness of fit and model checks. Finally some concluding remarks are given in Section 2.7.

2.2 LITERATURE REVIEW

There are several techniques available to model the number of road accidents. In earlier research, the relationship between road accidents and other variables was found by using a conventional multiple regression technique. A standard linear regression model was mostly used for modelling road accidents before the widespread availability of the Generalized Linear Model (GLM). Linear regression is based upon following assumptions:

- The response variable follows a normal or Gaussian distribution;
- The variance is constant over the observations in the model;
- The linear predictor is used directly to calculate the fitted values of the model; and
- The relationship between dependent variables and explanatory ones is linear.

The standard linear regression model is not appropriate when it is unreasonable to assume that data are normally distributed. Thus conventional linear regression models lack the distributional properties to describe adequately random, discrete, non-negative vehicle

accident events on the road as described by Maycock and Hall (1984), Jovanis and Chang (1986), Joshua and Garber (1990), and Miaou and Lum (1993) and many others.

2.2.1 Generalized Linear Model (GLM)

The theory of generalized linear models was first developed by Nelder and Wedderburn (1972). In these models the response variable is taken to be distributed according to a member of the exponential family of probability distributions. Members of this family include the Gaussian or normal, binomial, Poisson, gamma, inverse Gaussian, geometric, and negative binomial distributions. These models are based on a linear predictor which is a quantity calculated as a weighted linear combination of explanatory variables. It was found that by restructuring the relationship between the linear predictor and fitted values, non-linear relationships could be modelled. These models are known as generalized linear models (GLMs). This facilitates extension of classical linear models in respect of the various assumptions that all observations are independent or uncorrelated, the distribution followed is normal and the error term has constant variance. Nelder and Wedderburn (1972), Hilbe (1993), Francis (1993), and Green and Payne (1993) characterised generalized linear models by:

1. a random component for the responses, *y*, which has a distribution following the exponential family;
2. a systematic component expressed in the form of the linear predictor, $\eta = \mathbf{x}'\boldsymbol{\beta}$ and calculated from the product of the vector $\mathbf{x}$ of explanatory variables with the associated vector $\boldsymbol{\beta}$ of parameters to be estimated;
3. a known monotonic, one-to-one, differentiable link function $g(.)$ relating the linear predictor to fitted values.

According to this formulation, the generalized linear model can be expressed as:

$y_i = \mu_i + \varepsilon_i \qquad i = 1, ..., N$ *(McCullagh and Nelder, 1983, 26-27, ff)* 2-1

where μ_i is the expected value of observation *i* and is related to η_i by

$\eta_i = g(\mu_i),$

$g(\ )$ is the link function and ε_i is the random component.

The model describes η_i in the form of the linear predictor

$$\eta_i = \mathbf{x}_i^{'} \boldsymbol{\beta} \qquad \textit{(McCullagh and Nelder, 1983, 26-27, ff)} \qquad 2\text{-}2$$

where $\mathbf{x}_i$ is the vector of explanatory variables for observation *i*, and $\boldsymbol{\beta}$ is the associated vector of parameters.

In this model, the variance of the observations *y* is related to the mean μ by:

$$Var(y_i) = \phi V(\mu_i) \qquad i = 1, ..., N \qquad 2\text{-}3$$

where ϕ is the dispersion parameter and $V(\ )$ is a differentiable function called the variance function. The model follows a distribution from the exponential family such as normal, Poisson, binomial, negative binomial, exponential or gamma according to the nature of the data. The Poisson regression models possess most of the statistical properties desirable in describing road accidents. In Poisson generalised linear models, the log-linear model can be adopted for the relationship between explanatory variables and the Poisson mean parameter μ_i:

$$E(y_i) = \mu_i = \exp\left(\mathbf{x}_i^{'} \ \boldsymbol{\beta}\right) \qquad \textit{(Hilbe, 2007, 32, ff)} \qquad 2\text{-}4$$

where $\mathbf{x}_i$ is the vector of explanatory variables for observation *i,* and $\boldsymbol{\beta}$ is a vector of parameters.

The probability *P* and the likelihood ℓ functions are given as:

$$P(y \mid \mu) = \frac{e^{-\mu}\ \mu^{y}}{y!} \qquad (y \geq 0), \text{ and} \qquad \textit{(Hardin and Hilbe, 2001, 127-128, ff)} \qquad 2\text{-}5$$

$$\ell(\boldsymbol{\beta} \mid y) = P(y \mid \mu = \exp(\mathbf{x}' \ \boldsymbol{\beta}))$$

It is convenient to work with the logarithm L of the likelihood, and this is given for the mean u by

$$L(u;y)=\begin{bmatrix} -u_i = 0 & if\ \ y_i = 0 \\ \sum_{i=1}^{n}\{y_i \ln(u_i) - u_i - \ln \Gamma(y_i+1)\} & \end{bmatrix} \qquad 2\text{-}6$$

When the property of the Poisson distribution that restricts the variance to be equal to mean is not supported by the data, they are said to be either under-dispersed $\text{var}(y_i) < E(y_i)$ or as is usual for the road accidents data over-dispersed $\text{var}(y_i) > E(y_i)$. In this case, the standard errors of parameters estimated from a Poisson will be underestimated (Maher and Summersgill, 1996). The case of over-dispersion can be addressed by adopting the negative binomial distribution, which allows for variance to be proportional to the mean with a constant of proportionality exceeding unity.

The negative binomial model is derived by rewriting equation 2.4.

$$E(y_i) = \mu_i = Z_i \exp(\mathbf{x_i'}\ \boldsymbol{\beta}) \qquad 2\text{-}7$$

where Z_i is a gamma-distribution error term with mean 1 and variance α^{-1}. The parameter α corresponds to the over-dispersion parameter of a negative binomial distribution. The inclusion of this term allows variance of *y* to exceed its mean. Thus

$$\begin{aligned} &Z \sim Gam(\alpha), \\ &E(Z) = 1, \quad \text{var}(Z) = \alpha, \\ &E(y_i) = \mu_i \qquad \textit{(Hardin and Hilbe, 2001, 144-145, ff)} \qquad 2\text{-}8 \\ &\text{var}(y_i) = E(y_i)[1+\alpha\mu_i] \\ &\qquad = \mu_i + \alpha\mu_i^2 \end{aligned}$$

The negative binomial distribution has the form:

$$P(y \mid \alpha, \mu) = \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\,y!}\left[\frac{1}{1+\alpha\mu}\right]^{\frac{1}{\alpha}}\left[\frac{\alpha\mu}{1+\alpha\mu}\right]^{y} \qquad \textit{(Hilbe, 2007, 80, ff)} \qquad 2\text{-}9$$

where $\Gamma(.)$ is gamma function. The joint likelihood of α, μ is given by:

$$\ell(\alpha, \boldsymbol{\mu} \mid \mathbf{y}) = \prod_i P(y_i \mid \alpha, \mu)$$ *(Hardin and Hilbe, 2001, 146, ff)*

so that the joint log-likelihood of α , μ is

$$L(\alpha; \boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\alpha u_i}{1+\alpha u_i}\right) - \frac{1}{\alpha}\ln(1+\alpha u_u) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right\} \quad 2\text{-}10$$

When the data is over-dispersed, the estimated variance will be larger than the estimated mean. Due to this, the standard errors of the parameter estimates, which will be estimated appropriately, will be greater than those estimated from the corresponding Poisson model.

2.2.2 Generalized Estimation Equation (GEE)

GLMs are based on the assumption that the individual observations are mutually independent. This assumption is commonly known as *iid* (independent and identically distributed). In the case of repeated observations, correlated longitudinal or clustered data, this assumption is violated. In the present study of road accident data within Great Britain, the data have a panel structure with repeated observations: i.e. police force corresponds to a member of the panel, and each is measured repeatedly with time frames of days, month or years. Liang and Zeger (1986) introduced the Generalized Estimation Equation (GEE) to allow for correlated responses. GEE provides an extension of GLM in which the matrix of correlation between residuals of observations is generalized from its implicit diagonal form in GLM:

$$V(\mu_i) = \left[D\left(V(\mu_{it})\right)^{\frac{1}{2}} R(\psi)_{(n_i \times n_i)} D\left(V(\mu_{it})\right)^{\frac{1}{2}} \right]_{n_i \times n_i} \quad 2\text{-}11$$

(Hardin and Hilbe, 2003, 58, ff)

where $V(\mu_i)$ is a diagonal matrix and $R(\psi)$ denotes the within-panel correlation matrix. In the GLM model form, the within-panel correlation *R* is represented by the identity matrix.

There are several correlation structures that are commonly used including independent, exchangeable and autoregressive error structure. According to Hutchings (2003) the independent correlation structure is suitable when the number of observations per member of

the panel is small compared to number of members of the panel. The exchangeable correlation structure is used when it is assumed that correlation is constant between the observations. Autoregressive error structure is preferred when the observations have a natural order and as the time between the observations increase the correlation decreases. The details of some of the main correlation structures within GEE framework are described by Hardin and Hilbe (2003) as follows:

2.2.2.1 Independent structure

The independent structure that corresponds to GLM is defined as

$$R_{uv} = \begin{bmatrix} 1 & if\ u = v \\ 0 & otherwise \end{bmatrix} \qquad \textit{(Hardin and Hilbe, 2003, 59, ff)} \qquad \text{2-12}$$

2.2.2.2 Exchangeable structure

Exchangeable structure assumes a common correlation among observations within the panel. In this case ψ is scalar and the working correlation matrix has following structure:

$$R(\psi) = \begin{bmatrix} 1 & \psi & \psi \cdots\cdots\cdots \psi \\ \psi & 1 & \psi \cdots\cdots\cdots \psi \\ \psi & \psi & 1 \cdots\cdots\cdots \psi \\ \vdots & \vdots & \vdots \cdots\cdots\cdots\cdots \vdots \\ \psi & \psi & \psi \cdots\cdots\ 1 \end{bmatrix} \qquad \textit{(Hardin and Hilbe, 2003, 59, ff)} \qquad \text{2-13}$$

The GEE with an exchangeable correlation structure uses estimated Pearson residuals from fitting the model to estimate the common correlation parameter. The estimate of ψ using these residuals is $\hat{\psi} = \frac{1}{\hat{\varphi}} \sum_{i=1}^{n} \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \hat{r}_{iu}\ \hat{r}_{iv} - \sum_{u=1}^{n_i} \hat{r}_{iu}^{\,2}}{n_i (n_i - 1)} \right\}$ 2-14

where φ is the scale parameter and $\hat{r}_{it}$ is the estimated Pearson residual which is equal to:

$$\hat{r}_{it} = (y_{it} - \hat{\mu}_{it}) \Big/ \sqrt{V(\hat{\mu}_{it})} \qquad \text{2-15}$$

2.2.2.3 Autoregressive correlation of order 1 (AR1)

Autoregressive structure assumes time dependence for the association when observations of the members of a panel have a natural order. Autoregressive order 1 (AR1) weighs the correlation between two observations by their separation in time: as the difference in time between the observations increases the correlation decreases. In this case $\boldsymbol{\psi}$ is a vector and the correlation is estimated by using the Pearson residuals from fitting the model.

$$\hat{\boldsymbol{\psi}} = \frac{1}{\hat{\varphi}}\left[\sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{n_i-0}\hat{r}_{i,t}\,\hat{r}_{i,t+0}}{n_i},\ldots\ldots\ldots,\frac{\sum_{t=1}^{n_i-k}\hat{r}_{it}\,\hat{r}_{i,t+k}}{n_i}\right)\right] \qquad \text{2-16}$$

(Hardin and Hilbe, 2003, 66, ff)

2.2.2.4 Summary of the statistical methods

It is found that Generalized Linear Model (GLM) with Poisson distribution is a standard method used to model count response data. However, the Poisson distribution has equal mean and variance. Data that have greater variance than mean are termed as over-dispersed and negative binomial is the standard method used to model data that are over-dispersed relative to Poisson. Over-dispersion, which leads to larger residual deviance, can arise for several reasons, one of which is because some important explanatory variables have been omitted from the model. These may not even be available in the dataset. It can also arise because the process being modelled is fundamentally more variable than a Poisson process such as arises with the number of casualties when accidents occur according to a Poisson process. For a Poisson model, the expected value of residual deviance should approximately be equal to the residual degrees of freedom (McConway et al, 1999). In cases where the residual deviance of a Poisson model cannot be reduced to a value close to this, we consider adopting a negative binomial distribution instead to accommodate over-dispersion.

Furthermore, GLM structure does not accommodate the serial correlation which arises due to time series of data. In time series data the observations follow a natural ordering over time due to which successive observations are likely to exhibit correlation. Generalized Estimation Equation (GEE) can accommodate serial correlation in the data. In this study, the data used was for sequential days (Chapter 2, 4 and 5) and for sequential months (Chapter 3). So, use of the correlation structure of autoregressive order 1 (AR1) was investigated.

In the analysis of road accident data presented in this thesis, the results of the models using Poisson and negative binomial will be compared. GEE with AR1 is also used as it can accommodate the presence of serial correlation in the data. The results obtained by GEE and GLM will also be compared to identify any differences in the estimated parameter coefficients and their significance levels. This comparison will only be informal as both models (GLM and GEE) were fitted to the same data so that the estimates of the corresponding parameters are not mutually independent.

Further statistical methodology will be introduced as required in the course of this thesis.

2.2.3 Previous Studies

Various researchers have used linear regression, GLMs with Poisson and negative binomial distributions for modelling the road accident data. It was found from the previous studies that appropriate methods were not used in some of the studies to model count data. Bester (2001) used ordinary least squares linear regression without justification of its use for count data. Due to the unsuitability of this formulation, which admits negative estimates and has unsuitable error structure, the estimated coefficients and their significance may not be reliable.

Fridstrom et al (1995), Jones, Janssen, and Mannering (1991), and Greibe (2003) used generalized linear model (GLM) with log link function and Poisson error structure. However, in these studies they made no attempt to account the presence of over-dispersion in the data as suggested by Miaou and Lum (1993). Levine, Kim and Nitz (1995), Fridstrom et al (1995), and Memon (2006) used log-linear models with negative binomial error structure to accommodate over-dispersion in road accidents data for each day and month but did not discuss the presence of serial correlation and its effect on the modelling results. Due to this, the conclusions drawn from these studies may not be reliable.

Edwards (1996) used monthly number of road accidents and weather information recorded in STATS 19 data rather than independent meteorological data to identify some relationships in the eight regions of Great Britain. The author used linear regression for modelling the number of road accidents for each month. The presence of over-dispersion and serial correlation were not taken into account, so the conclusion drawn from this may not be reliable. Some of the

studies undertaken by various researchers using linear regression, and GLM with either Poisson or negative binomial are summarised as below:

Bester (2001) in South Africa developed a linear regression model to investigate the difference in road fatalities of individual countries. National infrastructure, transportation, and socio-economic variables from international databases were considered as explanatory variables. The final model included passenger car ownership, human development index (HDI), and the percentage of other vehicles as explanatory variables. It was found that numbers of fatalities are decreasing over time, which was ascribed to improvement in the physical and social infrastructure of those counties.

Miaou and Lum (1993) developed two conventional linear least-squares regression models and two log-linear Poisson regression models to investigate their ability to model vehicle accidents and highway geometric design relationships. They concluded that conventional linear models lack the distributional property to describe adequately random, discrete, non-negative, and typically sporadic vehicle accident events on the road. On other hand Poisson regression models possess most of the desirable statistical properties. However, if vehicle accident data are found to be significantly over-dispersed relative to their mean, then using the Poisson regression models may overstate the precision of estimates of vehicle accidents on the road. In that case, more general probability distributions have to be considered. This has led many authors to use log-linear negative binomial regression, which allows for dispersion at least as great as Poisson, with consequent reduction in stated precision of estimates (Maher and Summersgill, 1996).

Fridstrom et al (1995) used generalized linear Poisson regression models for each of the four greater Nordic countries. Monthly road accident counts for each county in the countries along with other databases which include gasoline sales, weather conditions, duration of daylight, changes in legislation and reporting routines, trend variable, variables for different counties and months were used. Three different models were estimated one for each of the number of injury accidents, number of fatal accidents, and number of users killed. LIMDEP 5.1 computer software was used with the maximum likelihood estimation method. It was found that exposure was the most important variable which explained 50 percent of systematic variation in fatal accidents and more than 70 percent in injury accidents.

Levine, Kim and Nitz (1995) analysed changes in daily motor vehicle accidents for the city and county of Honolulu. They found that road accidents occurring on each day fluctuate according to an interaction between traffic volume, weekday travel patterns, holidays, and weather. Beyond that, Fridays and particularly Saturdays have more daily accidents. Minor holidays generate more daily accidents, but major holidays generate fewer daily accidents primarily due to lower traffic volume. The combination of afternoon and rainfall was found to be particularly dangerous. High levels of unemployment appeared to reduce road accidents on each day.

Shankar, Mannering, and Barfield (1995) explored the frequency of occurrence of highway accidents on the basis of multivariate analysis of roadway geometrics (e.g. horizontal and vertical alignments), weather, and seasonal effects. The negative binomial model of accident frequencies is estimated. Models were estimated for accidents classified as sideswipes, rear end, parked vehicles, fixed objects and overturns. Interactions between weather and geometric variables were identified. It was proposed to avoid steep gradients and horizontal curves with low design speeds in areas with adverse weather.

Jones, Janssen, and Mannering (1991) developed a Poisson regression model for accident frequency in Seattle, USA. Six models each for one zone were developed to estimate the accident frequency and to identify characteristics peculiar to a specific day that might increase or decrease the expected number of road accidents. The seasonal effects, weekly trends, special events, and environmental factors were used as explanatory variables. Various conclusions were made and the results obtained were used for the development of the Seattle's accident management system.

SALIFU (2004) applied the generalized linear models framework for the development of negative binomial models of accident frequency for un-signalized urban junctions in Ghana. A total of 91 junctions were considered comprising 57 T-junctions and 34 crossroads with a total of 354 and 238 accidents for T and crossroads respectively obtained from the national accident database for the period 1996-1998. Traffic flow data was obtained by carrying traffic counts and spot speeds. Junction inventories were carried out to collect information about the site and geometry. Because of over-dispersion of the count data, negative binomial regression was used. The best models were found to be those based exclusively on traffic exposure functions (traffic flow) which explained 50 percent more of the systematic variation in

accidents at T-junctions than at crossroads. It was also found that T-junctions with yield control had a much lower accident rate than those with stop control.

Greibe (2003) developed a model for road accidents on urban roads in Denmark. He used accident, traffic flow, and road design data. Road accident data was collected from the official accident statistics database whereas traffic flow counts were collected from the municipality and converted to AADT counts. A total of 1,058 police recorded accidents were related to 314 road links. The GLM was used and the distribution of road accident counts was assumed to follow a Poisson distribution. Different models were developed for junctions and road links in urban areas. It was found that motor vehicle traffic flow was the most powerful variable in models for junctions whereas additional explanatory variables describing road environment, number of minor side roads, parking facilities, and speed limit proved to be significant and important variables for estimating the number of road accidents.

Abdel-Aty and Radwan (2000) used a negative binomial modelling technique for modelling the frequency of road accident occurrence in central Florida. The dataset that they analysed consisted of a total of 1,606 road accidents that occurred in the three years 1992-1994. It was found that heavy traffic volume, speeding, narrow lane width, larger number of lanes, urban roadway sections, narrow shoulder width and reduced mean width, increase the mean accident frequency. Different negative binomial models were developed based on the demographic characteristics of the drivers. It was also found that female drivers experience more road accidents than male drivers on roads that have heavy traffic volume, reduced mean width, narrow lane width, and larger number of lanes. Male drivers were found to be most involved in traffic accidents while speeding. The models also indicate that young and older drivers experience more accidents than intermediate aged drivers in heavy traffic volume, reduced shoulder, and reduced widths. Younger drivers have a greater tendency to be involved in road accidents while speeding or on roadway curves.

McCarthy (2002) developed a negative binomial regression models to analyse total, fatal and non-fatal injury alcohol-related crashes involving older drivers. He used data from the 58 counties of California for a period of 18 years (1981-1998) which consist of 1,044 observations. It was found that for the three categories: alcohol-related fatal crashes, alcohol-related non-fatal crashes, and alcohol-related total crashes, variance was greater than the mean, so that the negative binomial framework was preferred. The results indicated that risk exposure is a major determining factor, with the greatest effect on alcohol-related injury

crashes. Alcohol prices and income were also important variables. It was also found that speed limit policy rather than alcohol policies has the largest impact on alcohol-related crashes involving older drivers.

Lardon de Guevara, Washington, and Oh (2004) used negative binomial regression models to develop a planning level road accident prediction model for Tucson, Arizona. Separate models were developed for fatal, injury, and damage-only road accidents. It was found that population density, proportion of population aged 17 years or younger, and intersection density were significant variables for fatal crash models. However for injury and damage-only road accident models, population density, number of employees, intersection density, percentage of miles of principal arterial roads, percentage of miles of minor arterial roads and percentage of miles of urban collectors were significant variables.

Hall (1986) studied the personal injury traffic accidents that occurred at 177 four-arm single carriageway traffic signal junctions from urban areas of Great Britain from 1979 to 1982. Partial traffic flow data and pedestrian flow data was obtained from the Highway Authority; new counts were made at some junctions where this data was not available. The geometric data for each arm of the junction and the signal control characteristics were also incorporated into the models. The generalized linear modelling technique was used in GenStat software. It was assumed that the number of road accidents follows a Poisson distribution. Initial models were developed with only vehicle and pedestrian flows to which geometric, control and general factors were then added. Various conclusions were drawn about the influence of these characteristics on road accident frequencies.

Maycock and Hall (1984) used a generalized linear model with Poisson distribution to study roundabout accidents and to identify relationships between accident frequencies, traffic flow, and geometric design variables. The data sample included 84 four-arm roundabouts on main roads in the UK including small, conventional and dual carriageway roundabouts in both 30-40 and 50-70 mph speed limit zones. From the analysis of road accidents by accident type it was found that on small roundabouts accidents between entering and circulating vehicles were about 70 percent of the total whereas on conventional roundabouts the percentage is relatively evenly distributed between the accidents types of entering, circulating, approaching, single vehicle, other and pedestrian accidents. Different equations for each accident type were formed using GenStat and GLIM. The geometric variables considered for the model included entry path curvature, entry width, angle between arms, gradient, sight

distance, gradient, and approach curvature. Based on the values of the fitted coefficients, various conclusions were drawn about the effect of these variables on frequency of road accidents at roundabouts for each accident type.

Kulmala (1995) investigated factors that affect the road accidents at junctions outside urban areas in Finland. The accident data from 1983 to 1987 was used along with estimated traffic volumes. A total of 915 three and 847 four-arm junctions were considered. Generalized linear models with each of Poisson and negative binomial regression were used to estimate the number of casualties and to identify the most common accident class. The most important variables were found to be those describing the magnitude and distribution of motor vehicle volumes. Slight differences were observed in *t* values of parameter estimates between Poisson and negative binomial model. It was found that these models explained more than 80 percent of the expected systematic variation.

The literature review of the previous studies given in section 2.2.3 highlighted the various statistical techniques that have been used to model road accidents and casualties, and the explanatory variables that have been used for this. It was found that in various studies, linear regression and generalized linear models with Poisson regression were used in spite of having some shortcomings. Although Maycock and Hall (1984), and Hall (1986) used generalized linear models (GLM) with Poisson regression, they were aware of the presence of over-dispersion in the data. They addressed this by (a) scaling the standard errors of estimation and (b) offering procedures to estimate NB model.

Later, Miaou and Lum (1993), Levine, Kim and Nitz (1995), Fridstrom et al (1995) used generalized linear model (GLM) with negative binomial regression to accommodate the presence of over-dispersion in the response data. On consideration of explanatory variables, some that had not been used earlier were brought to attention for joint use with road accident data. In the studies carried out by Bester (2001), Fridstrom et al (1995), Levine, Kim and Nitz (1995), and Guevara, Washington, and Oh (2004) the variables of car ownership, time, gasoline sales, traffic flow, weather conditions, day of the week, major and minor holidays, proportion of population under 17 years or younger, percentage of the miles of different road classes and population density were used to identify their effect on the number of road accidents. In the present study an effort was made to use all the available information including this by joining the road accident data with other datasets.

2.3 DATA USED

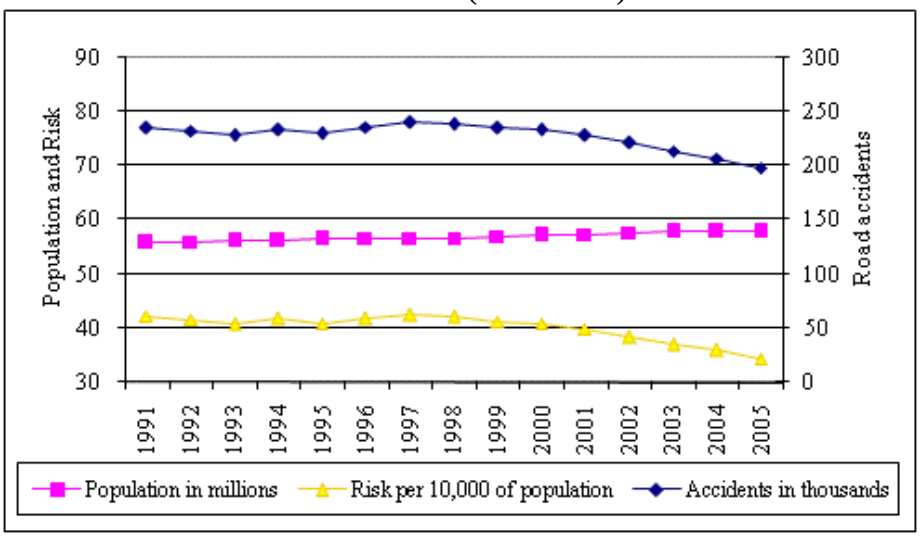
Road accident statistics in the Great Britain are compiled by the police. For each road accident that has caused personal injury, police authorities normally complete a STATS 19 form which provides details of road accident circumstances, information for each vehicle which was involved, and information of each person who was injured in the road accident. This whole dataset is maintained by the Department for Transport. For the present study the UK archive dataset was used which consists of total 3,417,878 road accidents recorded as occurring between 1st January 1991 and 31st December 2005. The required information for road accidents occurring on each day was extracted from the archive dataset using SPSS. As a result of this a new dataset was developed, containing information about all road accidents which occurred on each day from 1st January 1991 to 31st December 2005. Each day was given its original day name, month name, and year by using a calendar. Three separate variables were also included for each of all Public holidays, New-Year holiday, and Christmas holidays. The details of the days which are coded as Public holidays, New-Year holiday, and Christmas holidays are given in appendix Table A2.1. Two different datasets were prepared, respectively representing the whole of Great Britain and the 51 individual police forces, each of which corresponds to one or more local authority areas. Dataset 1 consists of 5,479 observations, each observation represents the number of road accidents on each day in Great Britain from 1991 to 2005. Yearly values of total distance travelled, population and number of registered vehicles was obtained from the Office for National Statistics and the Department for Transport. These variables were standardised to represent the character of transport activity rather than its scale. The details of this are given in section 2.5.1

Dataset 1 was further disaggregated to police force level to highlight the differences in number of road accidents across various locations. Dataset 2 consisted of 279,429 observations for the 51 police forces. Information of population, population density, number of registered vehicles and road length for each local authority were also obtained from the Office for National Statistics and the Department for Transport in addition to the above data. The values representing a local authority were then aggregated to police force level. The STATS 20 form which describes the instructions for completion of road accidents reports was used to aggregate the local authorities to police force level.

2.4 DATA ANALYSIS

The population, annual number of road accidents of Great Britain, and rate of involvement in a road traffic accident per 10,000 population is shown in Figure 2.1. It reveals that the population of Great Britain is slowly and continuously increasing. The estimated population for 2005 was 58.4 million. The figure also indicates that there was a slight change in pattern of population growth from 1996 to 1998 and again from 2002 to 2005; the lowest growth in population is observed during these two periods. On the other hand the annual number of road accidents after having slight fluctuations from 1991 to 1997 then followed a downward trend. It can be seen that 198,736 road accidents were recorded during 2005. The largest decrease of almost 24,000 accidents was observed during the three-years from 2003 to 2005. The risk rate per 10,000 population also decreased since 1997. The lowest rate was found for the year 2005 with a rate of 35 road accidents per 10,000 population.

Figure 2.1: Population, annual number of road accidents, and rate per 10,000 population of Great Britain (1991-2005)

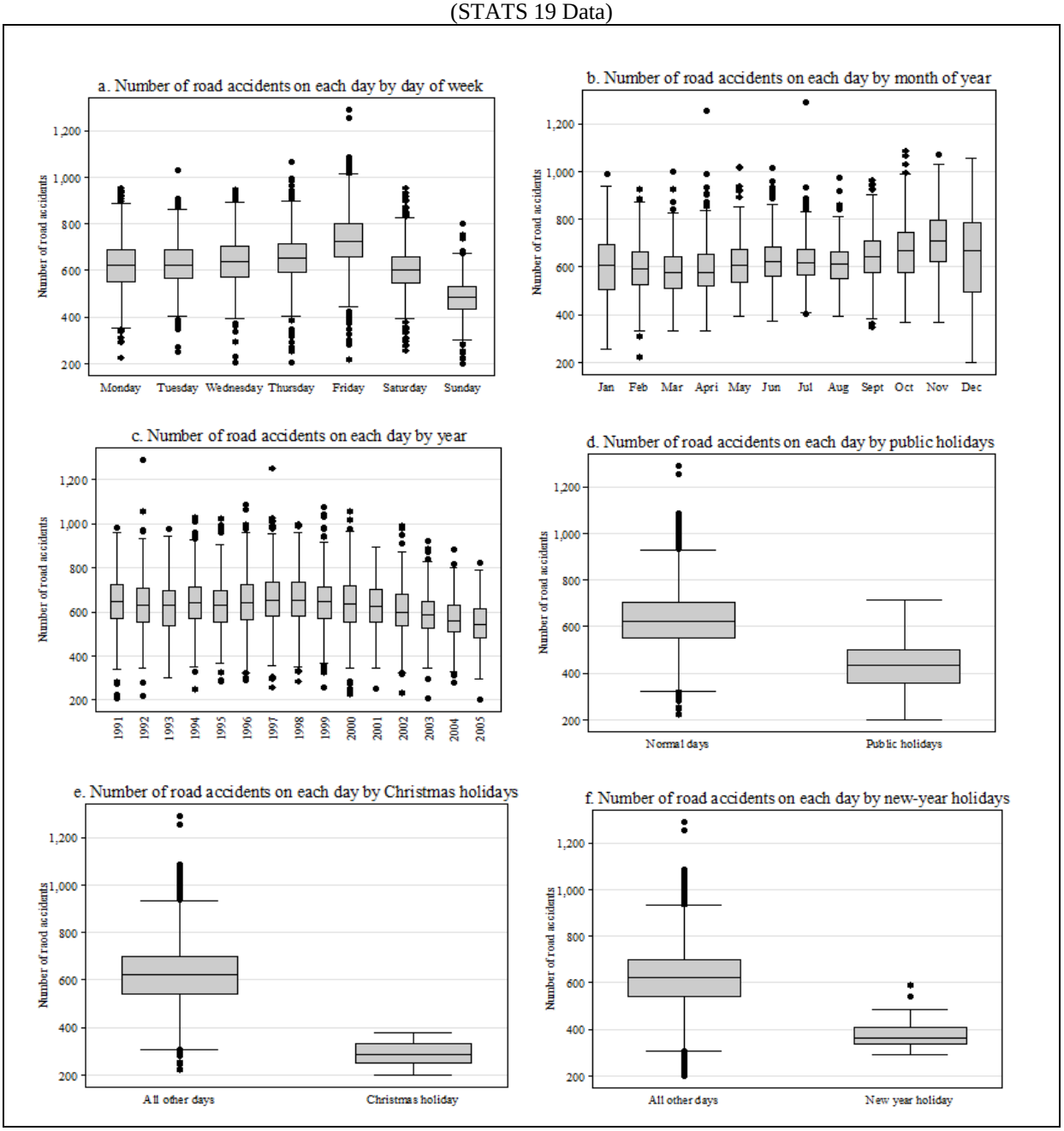


Source of data: Department for Transport (2011)

The detailed analysis of the dataset used for this study is shown in Figure 2.2. Each box plot in this figure consists of a central box which shows the inter-quartile range of data so that 50 percent of observations lie inside the central box. The horizontal bar within the box marks the median, upper and lower line of box represents the quartiles and whiskers indicate the minimum and maximum data values, unless outliers are present in the data. The whiskers extend to a maximum of 1.5 times of inter quartile range (IQR) from Q1 and Q3 beyond which other observations are considered outliers. If the median line with in the box is not in the centre than the data is said to be skewed. The circles in the box plot represent the outliers.

Figure 2.2: Box plots of road accidents in Dataset 1: 1991-2005

(STATS 19 Data)



Source of data: Department for Transport (2011)

The analysis of Dataset 1, which consists of road accidents for each day in the whole of Great Britain from 1991 to 2005, indicates the clear difference that is observed between weekdays and weekends. Comparatively higher number of road accidents occurs on Friday as it is the last working day of the week. Each day in the last 3 months of year (October, November and December) have more road accidents than others with each day in November having the highest road accidents. December and January are found to be more variable in terms of number of road accidents for each day than all other months as the IQR is greater for these

months which may be due to the number of Christmas and New-Year holidays. Christmas holidays have fewer road accidents than all other days, including all other holidays.

2.5 MODEL DEVELOPMENT

Regression models were developed for the number of road accidents occurring on each day with various combinations of explanatory variables having log-linear form, and each of Poisson and negative binomial error distributions by using the STATA software. In the first step, each model was developed with a constant term only and then a stepwise incremental approach was used to introduce different variables into the model. An offset variable was also used for which the details are given in section 2.5.3.

2.5.1 Variables used

The following variables from 1 to 14 were incorporated in the model for the national dataset (Dataset 1):

1. Day of the week (with 7 levels)
2. Month (12 levels)
3. Weekday 3 (with 3 levels: Weekday, Saturday and Sunday)
4. Season (with 4 levels: Spring, Summer, Autumn and Winter)
5. Day of week. Month interaction (84 levels)
6. Weekday 3. Season interaction (with 12 levels)
7. Time as a variate (measured in days, with values from 1 to 5479, corresponding to 1st January 1991 to 31st December 2005)
8. Public holidays (all bank holidays including Christmas and New-Year holidays)
9. Christmas holidays (25th December and associated holidays)
10. New-Year holidays (1st January and associated holidays)
11. Total distance travelled during the year (Vehicle kilometres)
12. Number of vehicles per head of the population
13. Distance travelled per head of population
14. Distance travelled per vehicle

Here the variable weekday 3 represents the difference between weekdays and two distinct weekend days. It has levels corresponding to Weekday, Saturday and Sunday. Similarly Season represents the difference between Spring, Summer, Autumn and Winter: Spring is from March to May, Summer is June to September, Autumn is October to November and Winter is from December to February.

In Dataset 1 annual figures of the total vehicle distance travelled, population and number of registered vehicles in Great Britain were used to derive variables 12 to 14, which are standardised so they characterise the transport activity rather than describe its scale. These circumstantial variables that represent characteristics of transport activity were preferred over the use of variables such as total vehicle distance travelled, population and number of registered vehicles in the interests of parsimony and to avoid inclusion of several variables in addition to the offset that describe the scale of transport activity, which would bring multicollinearity into the models.

For Dataset 2, which represents each police force, the same variables from 1 to 10 were used. All variables from 15 to 21 were specific to the police force and therefore characterise the area. The total number of registered vehicles in a police force that has a larger than average population would also be expected consequently to be larger than average, but use of vehicles per head of population provides a variable that characterises transport activity separately from its scale.

15. Population density (people per square kilometre)
16. Number of vehicles per head of population
17. Number of vehicles per kilometre of road length
18. Number of vehicles per square kilometre of surface area
19. Logarithm of population
20. Ratio of each road class to total road length
    a. LenTM (Length of trunk motorway)
    b. LenTR1 (Length of rural trunk road single carriageway)
    c. LenTR2 (Length of rural trunk road dual carriageway)
    d. LenTU1 (Length of urban trunk road single carriageway)
    e. LenTU2 (Length of urban trunk road dual carriageway)
    f. LenPM (Length of principal motorway)

g. LenPR1 (Length of principal rural single carriageway)
h. LenPR2 (Length of principal rural dual carriageway)
i. LenPU1 (Length of principal urban single carriageway)
j. LenPU2 (Length of principal urban dual carriageway)
k. LenBR (Length of rural B roads)
l. LenBU (Length of urban B roads)
m. LenCR (Length of total rural C roads)
n. LenCU (Length of urban C roads)
o. LenUR (Length of rural unclassified roads)
p. LenUU (Length of urban unclassified roads)

21. Police force as a factor (51 levels)

2.5.2 Coding systems for categorical variables in regression model

Categorical variables can be recorded into a series of variables for use in a regression model. There is variety of coding systems which can be used for coding categorical variables. A coding system reflects the comparison that is selected before running the regression models. Below are the coding structures that can be made in Stata software (UCLA, 2009):

- Simple coding
- Forward difference coding
- Backward difference coding
- Helmert coding
- Reverse Helmert coding
- Deviation coding
- Orthogonal polynomial coding

Deviation coding is preferred over others in this study as it reflects the deviations from the grand mean rather than the deviations from the reference category. In Stata, this can be achieved by using the DevCon directive which presents coefficients for factors from a statistical model in a way that achieves zero mean for their effects. When fitting a model, it is usual to set one coefficient to zero to avoid indeterminacy and hence to absorb that coefficient in the constant: usually this will lead to a non-zero mean. An adjustment can be calculated and applied to all factors (including any set to zero) so that they sum to zero; the

same adjustments can be accommodated in the constant so that the whole effect on the model is null. However, it was observed that the DevCon command showed reluctance to transform the coefficients correctly when interaction terms were added into the models. It is found that DevCon was suitable only for the main effects when there is only one reference category. Due to this, in Chapters 2, 4 and 5 where interaction variables are used, data was coded as combinations of 0,1 and -1 (deviation coding) as suggested by UCLA (2009) which resulted in coefficients for factors that had zero mean for their effects. In this case the results were verified by comparing the deviation coding (0, 1 and -1) and simple coding (1,0): as both produced the same estimated values (number of road accidents on each day) and log-likelihood results whereas deviation coding transformed the coefficients so that they refer to the group mean rather than a reference category. It is to be noted that in the case of unequal group sizes the intercept will represent the unweighted group mean rather than the grand mean. However, in chapter 3, this could be achieved by use of the DevCon command to transform the coefficients to have a zero sum as there were no interaction terms included in the model as explanatory variable.

2.5.3 Basic model structure

In this chapter, for all models that were developed for Dataset 1 and 2 as shown respectively in Figure 2.4 and 2.12, an offset variable was introduced. The offset variable represents the exposure to risk so that the risk per unit of exposure can be estimated directly from the linear predictor model. For this study, several variables were available for use as an offset, including vehicle distance travelled as vehicle kilometres, population, road length and number of registered vehicles. Road length is not preferred in this case as it cannot capture the temporal variations in the use of roads in an area. In the same way for number of registered vehicles it is difficult to capture the increased usage of vehicles (more distance travel) over time. Although Bird (2006) used road length and Fridstrom (1995) used fuel consumption as measures of exposure, it is difficult to determine where fuel is consumed which raises difficulties in the location of the exposure. The advantage of population over other measures of exposure is that in many cases the numbers are accurate and are available for specific groups of users. The vehicle distance travelled is probably the most often used exposure measure due to its availability at various levels of disaggregation. This can be related directly to the regional and temporal variations in road accident and casualty process.

The vehicle distance travelled and population is used as the main measure of exposure by IRTAD (2010) for comparison of road safety records in OECD countries.

In this study (Chapter 2) vehicle distance travelled on each day is used in the offset as a measure of exposure. The value of the offset variable is matched as closely as possible to the linear predictor for each unit of observation. This ensures that the linear predictor represents the risk as well as possible. Thus at the stage in model development when day of week was introduced as a factor into the linear predictor, the vehicle distance travelled was profiled according to the day of week by applying corresponding correction factors obtained from Department for Transport which account for the variations in vehicle distance travelled. Similarly at the stage when the month was introduced into the linear predictor the offset was profiled accordingly. Same process was repeated when weekday 3 and season were introduced into the linear predictor.

Beyond the offset variable, no others are used that describe the size of the unit of observation: to achieve this, other variables were coded in such a way as to characterise the unit of observation rather than to describe its scale.

Dataset 1:

The model used for Dataset 1 was

$$u_i = \exp\left(O_i + \mathbf{x}_i' \ \boldsymbol{\beta}\right) \qquad 2\text{-}17$$

where u_i is the estimated mean number of road accidents occurring on day *i*, and

O_i is the offset for day *i*

In this case $O_i = \ln\left(d_i\right)$

so that $u_i = d_i \exp\left(\mathbf{x}_i' \ \boldsymbol{\beta}\right)$ 2-18

where d_i is total distance travelled (vehicle kilometres) on day *i*.

This model structure then provides a direct estimate of risk r per unit of travel on day *i* as,

$$r_i = u_i / d_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}) \qquad 2\text{-}19$$

Dataset 2:

Dataset 2 is a disaggregated form of Dataset 1 which represents the 51 police forces of Great Britain. The aim of this disaggregation was to use the available information about these geographical areas which will ultimately increase the explanatory power of the model by identifying some systematic trends.

Information about the total distance travelled within each police force area was not available so attention was paid to other variables that could be used as measure of exposure as an offset in models. The following variables were considered and tested as an offset in Dataset 2 models.

1. Ln (total distance travelled nationally on each day)
2. Ln $\left[\left(\frac{Number\ of\ vehicles\ in\ Police\ force}{Total\ number\ of\ vehicles\ nationally}\right) \times National\ veh-km\right]$

Based on the experience obtained with dataset 1, initially national vehicle kilometres travelled on each day which was adjusted to take account of variation in distance travelled by day of week and month was used as an offset. This variable does not distinguish among police forces but it did at least allow for the different levels of usage over the day of week, month and years. The details of the modelling results are shown in appendix Tables A2.2.

After this, an adjustment was made in the vehicle kilometres which assumed that vehicle kilometres travelled within a police force area are directly proportional to the number of vehicles registered there. This offset variable distinguished among the police forces by taking account of variations in distance travelled by police force along with day of week and month variations. Based on the better *BIC* results and importance of these corrections to offset variable, it was considered and used as an offset in the models for Dataset 2. After this, the following model structure was used for Dataset 2 which will be discussed in later sections.

The mean number of road accidents occurring on day i in a police force j is estimated as

$$u_{ij} = \exp\left(O_{ij} + \mathbf{x}_{ij}' \boldsymbol{\beta}\right) \qquad 2\text{-}20$$

Then $u_{ij} = d_{ij} \exp\left(\mathbf{x}'_{ij}\ \boldsymbol{\beta}\right)$ 2-21

where d_{ij} is estimated total distance travelled (vehicle kilometres) on day i in police force j.

Following statistics were considered for the model preference, the definitions and formulas used are given as under;

2.5.4 Assessment of model performance

There are many ways to assess the performance of a statistical model. Each of these methods is informative but none of them is definitive. Rather, they can be used together to gain a balanced view of the performance of a model, and hence to guide selection of a preferred model.

The broad objectives of model development and selection followed in this study were to achieve a model that related variation in accident, number of vehicles involved and casualty numbers to explanatory variables in a way that represents a substantial proportion of the observed variation whilst respecting the nature of the variability of the data. The explanatory variables should have clear interpretation and they should have a good degree of mutual independence.

Various statistics including deviance residuals, log-likelihood values, information criteria, variance inflation factors and Durbin-Watson values which are described below in detail are used to guide the development and selection of a preferred model. During the modelling process, at various stages independent judgment and prior views on the importance of some explanatory variables was also used alongside the objective criterion.

An incremental approach was used to add variables into the model to observe their contribution to the performance of the models. As a starting point, a likelihood based objective measure (*BIC*) was compared for each of the models.

Analysis of temporal effects was also carried out to investigate the presence of any substantial systematic temporal effect that is not already represented in the model. Models in which this effect was established were not preferred. In order to identify the presence of

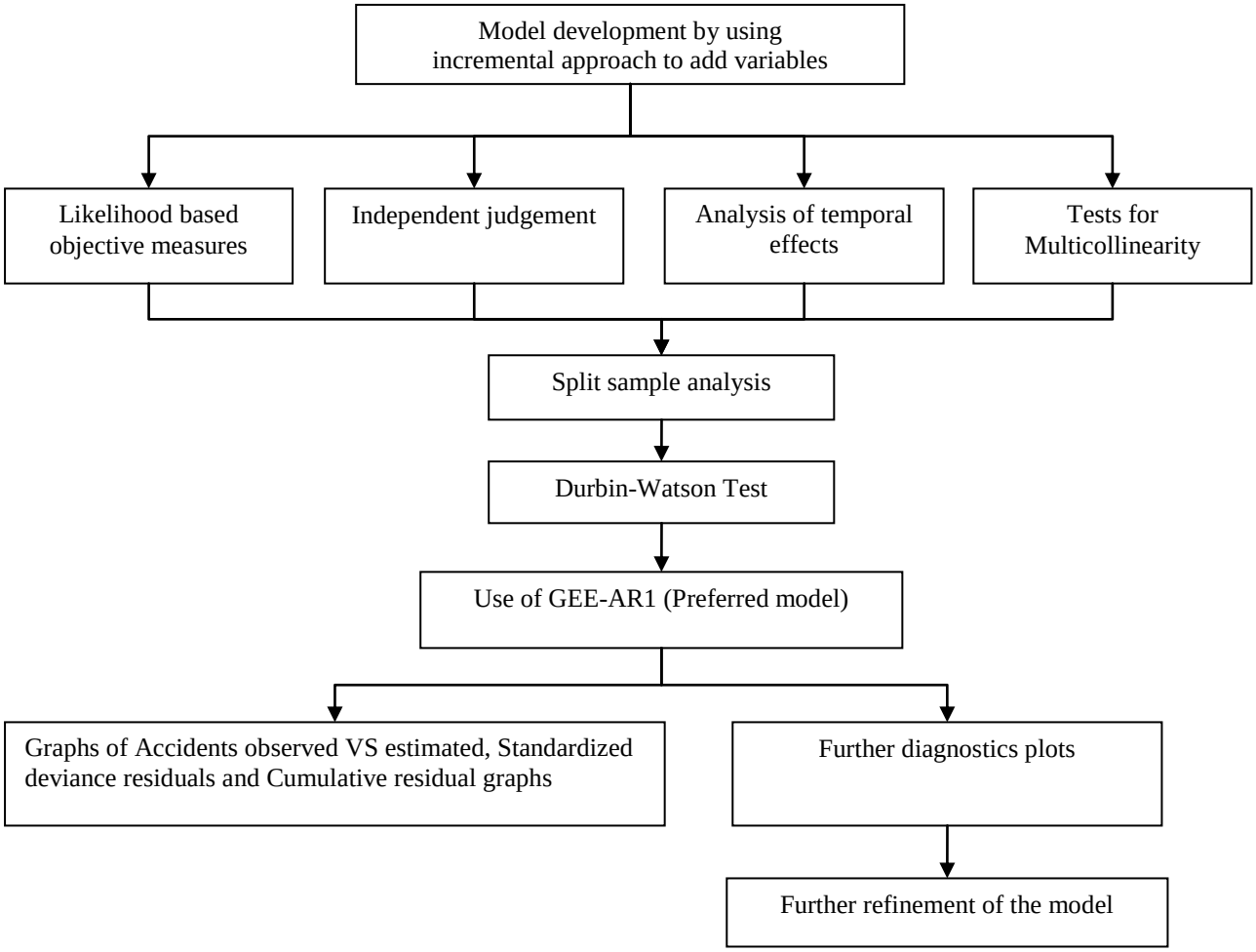
multicollinearity among the explanatory variables, variance inflation factors (VIF) were calculated. Attention was paid to the models where multicollinearity was observed among time and circumstantial variables and if found these models were not preferred as the estimated parameters will not necessarily represent their true effect. In cases where a high VIF value arose because of the structure of the data (for example, month and season), it was not taken as cause of concern.

After analysing these objective criterion, a model was taken forward out of the many developed in each case. In order to validate and check the consistency of the estimated parameters of the preferred model, split sample tests were carried out by dividing the whole dataset randomly in two portions. In order to check the consistency of model parameters the estimated coefficients of split sample were interchanged. Log-likelihood and deviance residuals values were estimated and compared. *T* test was also used to compare the estimated coefficients of the model by using these two portions of data to check parameters consistency and reliability.

The Durbin-Watson test was used to investigate the presence of serial correlation in the residuals. If serial correlation exists in the residuals then the GEE model formulation with AR1 error structure was used instead of GLM for the same set of variables. As the GLM and GEE models were fitted to the same dataset so the estimates of the corresponding parameters were not mutually independent: due to this only informal comparisons could be carried out to investigate the estimated parameters and their standard errors.

Apart from this, various other investigations were also undertaken which were used in conjunction with the tests mentioned above. These investigations include the graphs for the comparison of number of road accidents observed and estimated, standardised deviance residuals, cumulative residuals, deviance residuals against fitted values, normal quantile plot, scale location and Cook's distance plot were used for visual inspection to identify if any problem existed in the model. The Park test and Glejser test (Gujarati, 2009) are used to detect the presence of heteroscedasticity among the residuals along with some of the graphs mentioned above. If heteroscedasticity is found to be present then White's heteroscedasticity-robust standard errors are estimated by using the available procedure in STATA. Figure 2.3 shows the steps for the selection of the preferred model which are followed in all the chapters.

Figure 2.3: Steps in model selection procedure



In this section, measures of preference are discussed in detail while the model selection procedure, goodness of fit and model checks are disused in next section.

2.5.4.1 Likelihood and Deviance residual

The likelihood function presented in section 2.2.1 can be used to assess the goodness of fit of a model, and several further measures of model performance are based on it. It is to note that this assumes mutual independence of observations. In case the observations are not mutually independent, the likelihood will be overestimated. This will have the effect of exaggerating differences in log-likelihood and so will tend to favour elaborate models unduly.

Deviance provides an alternative to likelihood. The deviance is used as a measure of discrepancy of a generalized linear model; each unit *i* of observation contributes an amount

D_i as an increment to total deviance. For the Poisson model with observed number y_i and corresponding estimated number u_i, residual deviance is given by:

$$D_i = sign(y_i - \mu_i)\sqrt{d_i^2} \qquad \textit{(Hardin and Hilbe, 2001, 43, ff)} \qquad 2\text{-}22$$

where d_i^2 is the squared deviance residual which can be obtained according to the distribution as follows:

Poisson regression:

$$d_i^2 = \left\{ \begin{array}{ll} 2\,u_i & \text{if } y_i = 0 \\ 2\left\{ y_i \ln\left(\frac{y_i}{u_i}\right) - \left(y_i - u_i\right) \right\} & \text{otherwise.} \end{array} \right\} \qquad \textit{(Hardin and Hilbe, 2001, 230, ff)} \quad 2\text{-}23$$

Negative binomial regression

$$d_i^2 = \left\{ \begin{array}{ll} 2\ln\left(1+\alpha\, u_i\right)/\alpha & \text{if } y_i = 0 \\ 2y_i \ln\left(\frac{y_i}{u_i}\right) - \frac{2}{\alpha}\left(1+\alpha\, y_i\right)\ln\left(\frac{1+\alpha\, y_i}{1+\alpha\, u_i}\right) & \text{otherwise} \end{array} \right\} \qquad 2\text{-}24$$

(Hardin and Hilbe, 2001, 230, ff)

where α is the over-dispersion parameter.

The standardized residuals were obtained by multiplying the deviance residual D_i by the factor $\left(1-h_i\right)^{-\frac{1}{2}}$ where h_i is the leverage, which indicates the influence of observation *i*.

The total residual deviance *D* of the model is given by summation over all units:

$$D = \sum_{i=1}^{n} D_i \qquad 2\text{-}25$$

For Poisson, a properly fitted model the expected value of residual deviance should be approximately equal to the residual degrees of freedom (McConway et al, 1999).

2.5.4.2 Information Criteria

The maximised log-likelihood of a model will increase as further explanatory variables are introduced. This means that greater likelihood alone is not a suitable criterion for model selection. To address this, the Akaike Information Criteria (*AIC*) provides a likelihood-based measure of fit for a model that is adjusted according to the number of explanatory variables used:

$$AIC = -2L + 2k \qquad \textit{(Hardin and Hilbe, 2001, 45, ff)} \qquad 2\text{-}26$$

where L is the log-likelihood of the model and *k* is the number of explanatory variables.

This criterion can be used as an aid to model selection, with smaller values resulting in preferable models. Thus an elaboration to a model will be preferred if it increases log-likelihood by at least as much as the number of additional parameters in the model. In the case of dataset 2 which has 279,429 observations, use of an additional explanatory variable will be justified by an increase in likelihood of greater than 1.

However, larger datasets are more likely to justify the use of more explanatory variables. To address this, the Bayesian Information Criterion (*BIC*), which is also known as Schwarz Criterion (Schwarz, 1978), makes further adjustment according to the number of observations in the dataset:

$$BIC = -2L + k\ln(n) \qquad \textit{(Hardin and Hilbe, 2001, 45, ff)} \qquad 2\text{-}27$$

n shows the total number of observations in the dataset.

When this criterion is used, an elaboration to a model will be preferred if it increases the log-likelihood by at least $m.\ln(n)/2$, where *m* represents the additional degrees of freedom. In the case of dataset 2 which has 279,429 observations, an increase of 6.3 is required in the log-likelihood for one additional parameter in the model. This provides an alternative to the

Akaike Information Criterion that takes into account the number of observations, and so is well suited when large datasets are used. For this study the *BIC* is preferred over the *AIC* as it is more stringent and has a stricter entry requirement than *AIC* for additional parameters when large datasets are used. This helps to resolve over-fitting of models where many additional parameters are added to increase the likelihood, so *BIC* helps to promote a parsimonious model (Stata manual, 2001)

The log-likelihood values will not be reliable if the data observations are not mutually independent. Dependence in data structure occurs when the data observations are affected by common influences that are not represented in the model. In such a case the difference in the likelihood values, which is used in likelihood ratio test, will be overestimated. Due to this, likelihood values and all test based on them may not be reliable and hence are used cautiously in model selection.

Chandler and Bate (2007) proposed the adjusted likelihood ratio test for use when there is dependence in the data. However, in this study tests based on unadjusted likelihood values were used cautiously as the datasets were large and these tests were used as a guide in the model selection process along with other pertinent tests (see section 2.5.4) such as residual analysis, split sample test, graphs of observed and estimated values. In this way, model selection was carried out cautiously. However, in future it is recommended to adjust the log-likelihood values due to dependence and to identify the impacts on the likelihood, *BIC* and model selection process.

2.5.4.3 Likelihood ratio test

The likelihood ratio test can be used to compare the goodness of fit of two competing models that are nested. The model with additional variables was compared with the restricted model. The likelihood ratio statistic is:

$$X^2 = -2\left[L(\beta_R) - L(\beta_u)\right], \qquad \textit{(Chandler and Scott, 2011, 115, ff)} \qquad 2\text{-}28$$

where $L(\beta_R)$ is the log likelihood of the restricted model and $L(\beta_u)$ is the log likelihood of the unrestricted model. Under the null hypothesis that the restricted model is adequate, the

X^2 test statistic is χ^2 distributed with degrees of freedom equal to the difference in number of parameters between the restricted and unrestricted models (Washington, 2003).

2.5.4.4 Variance Inflation Factor

The variance inflation factor (VIF) is used to quantify multicollinearity among the explanatory variables. Stata estimated the values of VIF which can be used to adjust the standard errors of the parameter estimates, due to the presence of collinearity. A maximum acceptable value of 10 as proposed by Kutner (2004) is adopted in this study. The following formula is used in Stata to estimate the value of VIF.

$$VIF\left(\beta_j\right) = \frac{1}{\left(1 - R_j^2\right)} \qquad \textit{(Chatterji and Hadi, 2006, 236, ff)} \qquad 2\text{-}29$$

where j =1, 2, 3,……, k and R_j^2 is the multiple correlation coefficient of x_j on the other explanatory variables.

2.5.4.5 Durbin -Watson statistics

The Durbin-Watson statistic can be used to test the presence of first-order autocorrelation, and hence is used to analyse the residuals of a regression model. The test compares the residual for time period t with the residual from time period t-1 and develops a statistic that measures the significance of correlation between successive residuals. The formula for the statistic is:

$$d = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n}\left(e_t\right)^2} \qquad \textit{(Chandler and Scott, 2011, 66, ff)} \qquad 2\text{-}30$$

d = Durbin-Watson statistic

e = residual $\left(Y_t - y_t^e\right)$

t = time period counter

Table 2.2 shows regions of the acceptance and rejections of null hypothesis where d_l and d_u indicate the lower and upper critical values. The null hypothesis (H_0) is that there is no first order serial correlation among the residuals.

Table 2.2: Regions of acceptance and rejection of the null hypothesis at the α = 0.05 level for the presence of autocorrelation (Kendall and Ord, 1990, p268)

[0, d_l]	**[d_l , d_u]**	**[d_u , 4-d_u]**	**[4-d_u, 4-d_l]**	**[4-d_1, 4]**
Reject Null $H_{0:}$ **Positive** Autocorrelation	Neither accept nor reject	Accept the Null Hypothesis	Neither accept nor reject	Reject Null $H_{0:}$ **Negative** Autocorrelation

Significance points of d_1 and d_u at 95 percent significance level

	K=1		*K*=2		*K*=3		*K*=4		*K*=5	
n	d_1	d_u	d_1	d_u	d_1	d_u	d_1	d_u	d_1	d_u
50	1.5	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1.52	1.7	1.49	1.74	1.46	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
100+	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

K = number of independent variables in the equation

n= number of observations in the data

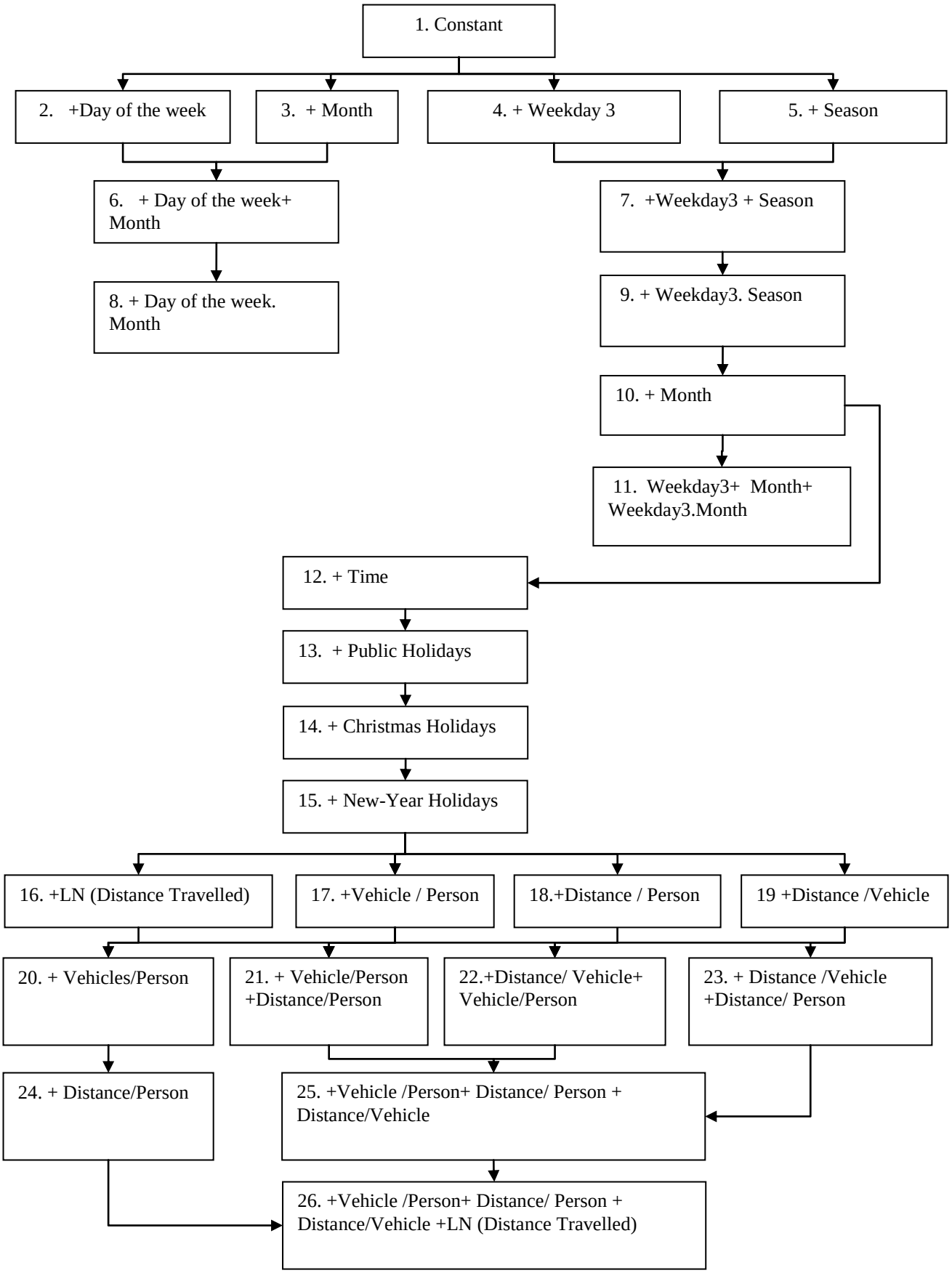
2.6 MODEL SELECTION PROCEDURE, GOODNESS OF FIT AND MODEL CHECKS

The model selection procedure as detailed in section 2.5.4 was applied to prefer the appropriate model out of the many available models. The results of all the developed models shown in Table 2.4 were compared; details are given in section 2.6.2.1. Section 2.6.2.2 to 2.6.2.5 shows the details of checks which were used to confirm that the most appropriate model has been preferred.

2.6.1 Model Selection Procedure

The procedure discussed in section 2.5.4 was followed to select the most appropriate model to represent the number of road accidents on each day. This can give some insights on the variables that are related to the number of road accidents and the nature of this relationship. Models were developed using Poisson and negative binomial regression as shown in Figure 2.4. The available variables were used in different combinations to observe their contribution to the performance of the models.

Figure 2.4: Lattice of model development for Dataset 1



The following section shows the results of the tests carried out for model selection:

1. *BIC* values were compared for all the models to assess their performance. Details of this are given in section 2.6.2.1
2. The models were analysed primarily with the intention to investigate that there is no substantial temporal effect remaining. Details of this are given in section 2.6.2.2
3. Variance inflation factors were calculated to check for the presence of multicollinearity among the explanatory variables. Details of this are given in section 2.6.2.3
4. Split sample tests were carried out to validate the performance of the preferred model by cross-comparing the coefficients, deviance and log-likelihood values. Details of this are given in section 2.6.2.4
5. The Durbin-Watson test was used to detect the presence of serial correlation in the model residuals. Details of this are given in section 2.6.2.5.

2.6.2 Model selection process, goodness of fit and model checks for Dataset 1

This section shows results of the tests discussed above to select the preferred model. The goodness of fit of the preferred model and various other checks as described above were applied to validate the model are shown in detail as below:

2.6.2.1 Poisson and negative binomial regression model for Great Britain (Dataset 1)

2.6.2.1.1 Poisson regression model

The model development started with Poisson regression modelling using the log link function. The ultimate aim was to establish the relationship between road accident numbers occurring on each day and the explanatory variables from 1-14 as shown in section 2.5.1. The quality of model fit was assessed according to the Bayesian Information Criteria (*BIC*). A total of 26 models were developed with different combinations of variables as shown in Figure 2.4. The logarithm value of the total distance travelled on each day was used as the offset with all of these models and this was profiled where possible to correspond to the explanatory part of the associated model. In particular, for models in which the day of week and month was used, the offset was adjusted to take account of the associated variations in distance travelled using the correction factors obtained from the Department for Transport

which are shown in appendix table A2.3 and A2.4. The effect of applying these corrections is that the estimated coefficients represent the direct risk per unit of distance travelled.

In the process of model development, the day of week and month corrections to the offset were applied only when the corresponding variables were introduced into the model. In model 2, the offset was only adjusted for day of week corrections as it has only day of week as an explanatory variable. In the same way in model 3, the offset was adjusted by only using month correction factors. However in model 6, day of week and month corrections were applied together as this model has both (day of week and month) as explanatory variables. The cases where simplified categorical variables such as weekday 3 and season were used, as in model 4 and 5, the profile of day of week and month of year adjustments to the offset were retained. From model 6 onwards, both day of week and month corrections were applied together to the offset in all models. Table 2.3 shows the list of models and the corrections applied to the offset.

Table 2.3: Details of the correction applied to the offset in models

Corrections factors used with the offset variable			
Model No.	Corrections applied to offset	Model No.	Corrections applied to offset
1	None	4	DoW
2	DoW	5	Month
3	Month	6-26	DoW and Month

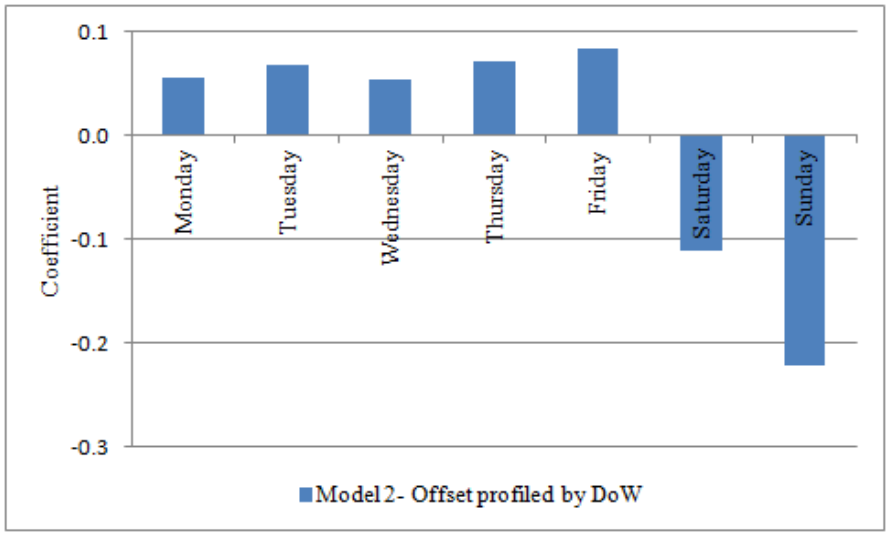
DoW represents the Day of week

Model 1 was developed by using constant term only in which no adjustments were applied to the offset variable to adjust the distance travelled by day of week. This model gave *BIC* of 212,488. A stepwise approach was then used for introducing explanatory variables. An improvement in the value of *BIC* of about 42,000 was observed after introducing the day of the week variable with 6 degrees of freedom into the model: this improved the *BIC* to 170,516.

In model 2, day of week was introduced and offset adjusted accordingly, it was found that all weekdays (Monday-Friday) had similar coefficients, showing that the risk per unit of travel is

similar. By contrast, Saturday and Sunday had substantially different values of risk. This led to the introduction of a new variable weekday 3 in model 4 with only three variables representing Weekday (i.e. any of Monday-Friday), Saturday and Sunday. The *BIC* of model 2 was better by value of 149 than model 4 suggesting that day of week performed better than weekday 3 when used individually. Figure 2.5 shows the coefficients of day of week form model 2 when the offset was profiled by day of week corrections to take account of variations in distance travelled.

Figure 2.5: Coefficients of Day of week from model 2 (Dataset 1)



In model 3, month variable was introduced and offset variable was adjusted only to take account of variations in distance travelled by month. This gave *BIC* value of 217,207 which was not better than model 2 where day of week variable was used. Model 4 with weekday 3 variable which was simple version of model 2 produced better results than model 3.

In model 5, seasons of year (Spring, Summer, Autumn and Winter) were introduced in the explanatory variable. This further led to the development of model 6, 7, 8 and 9 where day of week and month, weekday 3 and season, day of week, month and their interaction terms, and weekday 3, season and their interaction terms were used respectively. Model 9 was the simplified version of model 8 with 72 fewer degrees of freedom. As we understand that the number of road accidents also varies by month which is evident from the estimated coefficients of model 3, due to which month was included in model 10 along with weekday 3, season and their interaction. This model helps to capture the variability in the number of road accidents in addition to the season variable already in the model. Further to this, in model 11

weekday 3, month and their interaction variables were used to identify the improvement in *BIC* of the model in comparison to model 8 and 10.

By comparing the results of model 8 (day of week, month and their interaction), model 10 (weekday 3, season, their interaction and month) and model 11 (weekday 3, month and their interaction) it was found that model 8 had better *BIC* values than other two models, but it has 84 degrees of freedom. Out of these 3 models, model 10 was carried forward based on our own judgement and its performance in terms of *BIC* when using negative binomial regression where it performed better than model 8 and 11 (see Table 2.4). Model 10 has 63 fewer degrees of freedom than model 8 and explanatory variables of weekday 3, season, interaction of weekday 3 and season, and month.

Model 10 has *BIC* value of 173,599. An improvement of about 39,000 was observed in the value of *BIC* for model 12 in comparison to model 10 when the Time variable was included. This established the presence of temporal trend. Gradual improvement in the value of *BIC* was observed as further variables were included in the model. Each of the variables of Public holidays, Christmas holidays and New-year holidays in models 13-15 improved the *BIC* value by 9,000, 2,445 and 330 respectively.

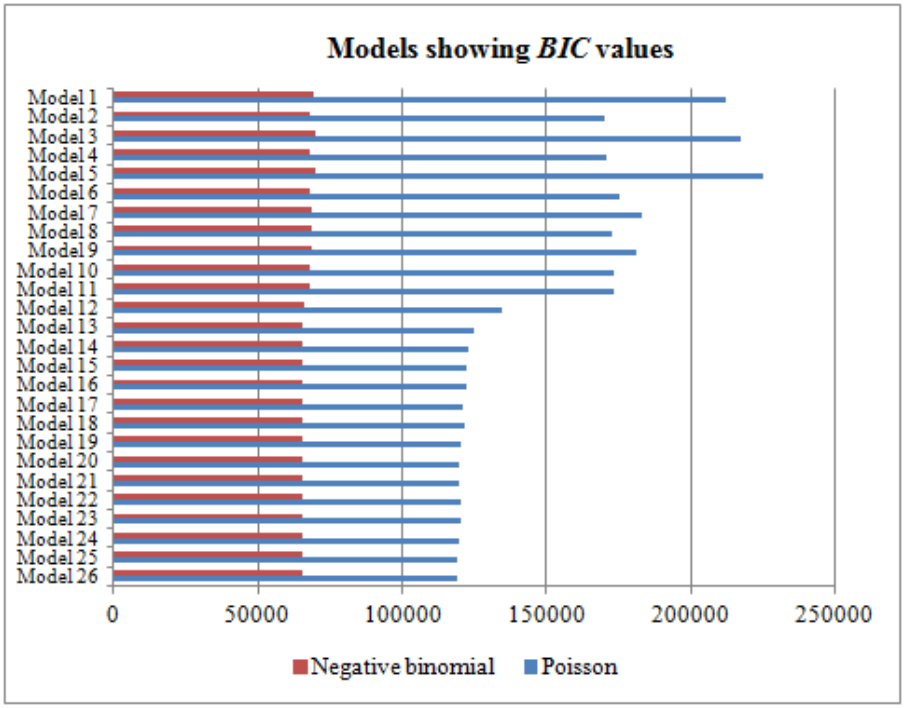
In model 16, the logarithm of the annual distance travelled was introduced as an explanatory variable to investigate whether it had an effect beyond the linear one that is represented through offset. This was evaluated by the change in the *BIC* value of the model. The addition of this variable resulted in improvement in comparison to model 15 of only 373, which is small in comparison to the contribution from other variables. This shows that any non-linear effect of the distance travelled is not strong. Due to this, it is represented only in the offset.

In model 17-19, the variables of vehicles per person, distance travelled per person and distance travelled per vehicles were used individually. The *BIC* of model 19 with variable of distance travelled per vehicle is better by value of 577 and 1,713 than model 17 and 18 where vehicles per person and distance travelled per person were used respectively.

After adding all the variables in various combinations, as shown in Figure 2.4 model 26 with 29 degrees of freedom had better values of *BIC* than any of the other models. The value of *BIC* for model 26 was 119,029. After including weekday 3, season, interaction of weekday 3

and four season, month, time, Public holidays, Christmas holidays, New-Year holidays, total annual distance travelled, vehicles per person, distance travelled per person and distance travelled per vehicle the mean deviance residual for the final model was still 13.50 which showed that the model still leaves a substantial amount of unexplained variability. The results of all 26 models are shown in Table 2.4. The graph showing the performance of the models in terms of *BIC* is also shown in the Figure 2.6.

Figure 2.6: Comparison of the *BIC* values of the models (Dataset 1)



2.6.2.1.2 Negative binomial regression model

Due to the substantial amount of variability in the data, negative binomial regression was carried out. In Stata software the value of the over-dispersion parameter α is not estimated by the glm command so that the nbreg command was used initially to estimate it. Hilbe (2007) noted that when α is significantly different from zero, then a negative binomial model is preferred to a Poisson one. This estimated value of α is then used with the Stata glm command to estimate the remaining model parameters. Although the model parameters and standard errors produced by both commands were same, the glm command was used in order to take advantage of other statistical diagnostics that are available in Stata software to evaluate the model fit (Hilbe, 2001).

The same procedure as used in section 2.6.2.1.1 was carried out by making incremental changes into the model. All 26 models shown in Figure 2.4 were developed. The *BIC* values were used to compare efficiency and effectiveness of models. The ultimate aim was to establish informative models to which the explanatory variables contribute. This was achieved by investigating the effects of introducing the explanatory variables and by analysing the model residuals. It is found that estimated value of the over-dispersion parameter α of the negative binomial distribution is statistically greater than zero in each of the models hence justifying the use of negative binomial regression.

The same procedure was applied to adjust the distance travelled by day of week and month as explained in section 2.6.2.1.1 and Table 2.3. The first model was developed by using a constant term only, which gave the *BIC* value of 69,514. Better *BIC* values were obtained when the day of the week variable was added into the model as an explanatory variable at the same time as day of week correction applied to offset variable to account for the variations in distance travelled by day of week. Use of the simplified variable weekday 3 (Weekday, Saturday, and Sunday) in model 4 resulted in better *BIC* than day of week in model 2 which suggest that weekday 3 variable (with 3 levels) has performed better than day of week variable (with 7 levels) when negative binomial regression is used. On the other hand the use of month with 12 levels (model 3) with *BIC* value 69,765 performed better than Season with 4 levels (model 5) with *BIC* value of 69,939 showing that use of month variable is justified.

From model 6 onwards different variables were used in combinations in the linear predictor. In model 6, day of week and month variable were used together, while in model 7 the simplified variables of weekday 3 and season were used. Model 7 did not perform better than model 6. By comparing the *BIC* values for model 6 and 7 it was found that *BIC* of model 6 was better by value of 216. Greater improvements were obtained when the respective interaction variables was introduced in model 8 and 9. For model 8 the value of *BIC* was found to be 68,631 with 84 degrees of freedom when the day of week, month and their interaction variable were used together while model 9 has slightly better *BIC* (by 251) value than model 8 and fewer degrees of freedom being associated with the simplified interaction variable. The *BIC* value of model 9 was 68,380, this shows that *BIC* supports use of the more parsimonious model. Month variable was also introduced in model 10 to account for extra variation available in data which was evident in model 3 where month variable performed better than seasons (Model 5). This further addition of month as explanatory variable

improved the *BIC* of model by 239 in comparison to model 9 which justifies the use of month variable in model 10.

In model 8 (day of week, month and their interaction) and in model 10 (weekday 3, season, their interaction and month) were introduced. As it was found earlier in this section that weekday 3 and month variable performed better individually than day of week and season respectively, due to this they were used together in model 11 along with their interaction variables and compared to model 8 and 10. It was observed that the *BIC* value of model 10 was better by 490 and 115 than model 8 and 11 respectively which justifies the preference for model 10 in comparison to model 8 and 11. Due to this, model 10 was considered further by adding other available explanatory variables.

A large improvement of about 1,900 in *BIC* was observed when the Time variable was added in model 12. After this the Public holidays, Christmas holidays and New-year holidays were added incrementally which resulted in improvement of 632, 223 and 22 in model 13, 14 and 15 respectively. The *BIC* of model 15 was found to be 65,350.

After this the logarithm values of the annual distance travelled were introduced into the explanatory part of model 16 to investigate the improvement in model performance. The addition of this variable resulted in improvement in *BIC* of only 27 in comparison to model 15. So, non linear effect of distance travelled is not strong and this variable will be represented only in the offset. After this, circumstantial variables of vehicles per person, distance travelled per person and distance travelled per vehicle were used individually in model 17-19. It was found that model 19 with distance travelled per vehicle had better *BIC* value of 65,190 than model 17 and 18 where vehicles per person and distance travelled per person were used respectively.

From model 20 onwards these circumstantial variables were used in various combinations into models which further improved the *BIC.* After including all variables that were available, the values of *BIC* improved to 65,122 for model 26. This resulted in an improvement of 4,392 (about 6 percent) in comparison to model 1 by adding weekday 3, season, interaction of weekday 3 and season, month, time, public holidays, Christmas holidays, new-year holidays, logarithm of annual distance travelled, vehicle per person, distance travelled per person and distance travelled per vehicle.

Table 2.4: Results of all models for the whole of Great Britain (Dataset 1)

Results of model for the whole of Great Britain (Dataset 1)							
		Poisson Distribution			Negative binomial		
Model	**D.F**	**MD**	**L.L**	***BIC***	α	**Likelihood**	***BIC***
1	1	30.5	-106,240	212,488	0.04816	-34,753	69,514
2	7	22.9	-85,228	170,516	0.03529	-33,922	67,905
3	12	31.4	-108,552	217,207	0.04957	-34,831	69,765
4	3	22.9	-85,319	170,665	0.03535	-33,927	67,879
5	4	32.8	-112,444	224,922	0.05185	-34,952	69,939
6	18	23.8	-87,679	175,512	0.03676	-34,031	68,217
7	6	25.2	-91,605	183,261	0.03904	-34,191	68,433
8	84	23.5	-86,126	172,975	0.03571	-33,954	68,631
9	12	24.9	-90,551	181,204	0.03829	-34,139	68,380
10	21	23.5	-86,709	173,599	0.03606	-33,980	68,141
11	36	23.5	-86,569	173,448	0.03597	-33,973	68,256
12	22	16.3	-67,156	134,501	0.02496	-33,019	66,227
13	23	14.6	-62,496	125,191	0.02204	-32,699	65,595
14	24	14.2	-61,270	122,746	0.02109	-32,583	65,372
15	25	14.1	-61,100	122,416	0.02096	-32,567	65,350
16	26	14.0	-60,910	122,043	0.02082	-32,549	65,323
17	26	13.8	-60,247	120,718	0.02048	-32,507	65,238
18	26	14.0	-60,815	121,854	0.02076	-32,542	65,308
19	26	13.7	-59,959	120,141	0.02028	-32,483	65,190
20	27	13.7	-59,838	119,908	0.02021	-32,474	65,180
21	27	13.6	-59,802	119,837	0.02019	-32,471	65,174
22	27	13.7	-59,944	120,121	0.02028	-32,482	65,197
23	27	13.7	-59,958	120,149	0.02028	-32,483	65,198
24	28	13.6	-59,730	119,702	0.02015	-32,466	65,174
25	28	21.7	-59,450	119,142	0.01997	-32,443	65,127
26	29	13.5	-59,389	119,029	0.01992	-32,436	65,122

**D.F = Degrees of freedom, M.D = Mean Deviance, L.L= log-Likelihood values, BIC= Bayesian information criterion*

The comparison of *BIC* values of negative binomial and Poisson for all the models is shown in Figure 2.6. Detailed results of all the models are shown in Table 2.4. This shows that the negative binomial fitted consistently better than the Poisson, due to its accommodation of over-dispersion. Improvements in the fit of the negative binomial model were smaller than for the Poisson because this corresponded to making explicit dependence of some part of the

dispersion. Due to this, we preferred negative binomial in comparison to Poisson regression models

From the modelling results shown above it was observed that variations in risk per unit of distance travel within week are represented adequately by weekday 3: so it was concluded that risk per vehicle kilometres is roughly equal among weekdays, but substantially different on each of the Saturday and Sunday. However, the use of month (12 levels) is justified in presence of the simplified variable of season (4 level). Systematic variations in risk among days of week over the month in the year were found to be represented adequately by the interaction of weekday 3 and season. This has the advantage of parsimony over interaction between weekday 3 and month because it requires only 6 additional degrees of freedom rather than 22 or more for other formulations.

2.6.2.2 Analysing the temporal effects

In this section the negative binomial models fitted in section 2.6.2.1.2 were analysed further to investigate whether there was any substantial systematic temporal effect that was not represented in the model. This was carried out by adding time and the square of time variables to the models. The resulting improvement in *BIC*, coefficients and *t* values of time and square of time, and their variance inflation factors (VIF) were examined. Because models 1-11 do not include time variable, both time and square of time variables were added to those models. From model 12 onwards when time variable was already present only square of time was added to investigate the presence of substantial quadratic temporal effect that was not represented by other explanatory variables.

From the results shown in appendix Table A2.5, substantial improvements in the value of *BIC* were observed for model 1-11 (except model 1, 2 and 4) when time and square of time variables were added to the models. In each of these models the *t* values of time was found to be non-significant whereas square of time had significant *t* values.

From model 12 to 15 only square of time variable was added as time was already included in the models. This resulted in improvement of *BIC* which was comparatively smaller than the initial models (model 1-11).

From model 17 onwards when circumstantial variables were included in different combinations, introducing the square of time resulted in smaller improvement in *BIC* whilst VIF increased showing correlation between time and circumstantial variables. In model 19 (with distance travelled per vehicle) the *t* value of time and square of time was 2.43 and -8.81. However, the *BIC* improved by only 69 and the estimated value of VIF for these variables was 23 and 77, which is high showing that some of the circumstantial variables in the model had non-linear temporal trend. The most detailed model 26 which had better *BIC* value than all other models showed there is only improvement of 2 in the *BIC* value when square of time variable is incorporated into the model. This small improvement shows that quadratic temporal trend in the data has adequately been represented by other variables in this model.

2.6.2.3 Checking for presence of multicollinearity

Multicollinearity can arise in the data due to associations among the explanatory variables. A consequence of its presence is that some statistical inferences about the data may not be reliable (Washington, 2003). Multicollinearity can cause some of the following problems in the results estimated by models:

- The standard errors of parameter estimates are likely to be high;
- The magnitude and sign of the parameter estimates are unreliable and can change from one sample to another.

As a result of this, the validity of inferences drawn from the model will be undermined. In this study it is evident from the structure of the data that some explanatory variables such as month and season are correlated. Keeping this point in mind, focus was on the circumstantial variables that had collinearity with time. In order to investigate the multicollinearity, variance inflation factors were estimated using formula 2-29 as used by the Stata software. These values of VIF can be used to estimate any consequent increase in the standard errors of the coefficient estimates.

Table 2.5 shows the individual VIF of variables used in model 16-26. The VIF of the variables used in initial models (1-15) are not presented as it is understood that there will correlation due to association among the variables (interaction variables, month and seasons), hence it was not considered to be a cause of great concern.

From model 16 onwards (except model 19) the VIF for time and circumstantial variables are high in most cases. The explanatory variables of time, distance travelled, vehicles per person, distance travelled per person and distance travelled per vehicle had high VIF because these quantities have established trends in their development over time. As a consequence, the partial effects of these circumstantial variables can not be estimated reliably when more than one of them appears in the same model.

In model 19 each of time and distance travelled per vehicle had acceptable VIF of 6. This suggests that there is no strong trend in distance travelled per vehicle over time. Other circumstantial variables when used together produced better *BIC* results but had multicollinearity. Because of this the effects of these variables can not be identified correctly as these quantities have established trends over time, due to which they were not preferred. As a result we look for a model that has just one of these circumstantial variables. Due to this, model 19 and its output values will be further analysed in the following sections.

Table 2.5: Variance inflation factors of variables for Dataset 1

Model	Time	Ln(D.T)	V/P	D/P	D/V
16	54.2	53.9			
17	42		42.2		
18	32.1			31.9	
19	6.1				6.1
20	55.9		43.5		
21	65.6		42.6	32.4	
22	64.1		114.3		16.6
23	65.9			42.5	8.1
24	279.2	6,344	67.1	3678	
25	65.99		2,696	1,002	514
26	549.2	13,682	6,622	4,047	1,108

Ln(D.T)= logarthim of distance travelled, V/P= vehicles per person, D/P= distance travelled per person, D/V= distance travalled per vehicle

2.6.2.4 Split sample tests

After analysing the *BIC*, temporal effects and *VIF* values according to the criteria discussed in section 2.5.4, model 19 was taken forward for further investigation. The detailed reasons of perference of model are given in section 2.6.2.6.

In order to validate and check the consistency of this model and its parameters estimates, split sample validation tests were carried out by dividing the whole sample randomly into two portions. The following procedure was adopted to achieve a 50-50 split. A uniform random variate in (0,1) was generated for each record and the whole dataset was then sorted using the random number. The first 50 percent of the observations (2,739) were used as Dataset B whereas remaining 50 percent of observations were considered as Dataset C. The following datasets were used to cross-check and validate the results of model 19.

Full dataset = Dataset A
Dataset first portion = Dataset B
Dataset second portion = Dataset C

Stata was used to estimate the parameters of model 19 with negative binomial error distribution for each of the Datasets B and C. In order to check the consistency and reliability of the model parameters, the coefficients estimated from Dataset B were used with Dataset C to estimate the number of road accidents on each day and after that values of log-likelihood and total deviance were estimated using equations 2-10 and 2-22 respectively. The corresponding process was repeated using coefficients estimated from Dataset C with Dataset B. After this, in order to further check the consistency of the estimated parameters the coefficients from dataset B and C were compared by using the *T* test.

The results in Table 2.6 show that values of the log-likelihood and total devaince are consistent and almost same for the Datasets B and C. For Dataset B the log-likelihood value estimated was -16,244 whereas for Dataset C it was found to be -16,229. Interchanging coefficeints between Datasets B and C produced only a small change in the values of log-likelihood and total deviance, making these values slightly less perferable than the initial values. The coefficients of dataset C when used with dataset B produced the log-likelihood of -16,265 which had the difference of only 21 from the value optimised for that dataset. Because the model parameters are not optimised in this case, there are 25 more degrees of freedom in the residuals: this gives rise to a likelihood ratio test of 42 on 25 degrees of freedom, which is less than the critical value of 44.31 at 0.01 significance level. Therefore the null hypothesis can not be rejected that parameters fitted to dataset C are as appropriate for dataset B as these fitted to that dataset. In the same way when coefficients of dataet B were used with dataet C that also produced the difference of 22: this gives rise to a likelihood

ratio test of 44, as a result of this null hypothesis can not be rejected at 0.01 significance level that parameters fitted to dataset B are as appropriate for dataset C. In general it is found that values of log-likelihood -32,483 and devaince 5,502 for Dataset A are better than the summation of other two models. This further confirms that the parameters of the model are consistent and reliable.

After this the coeffiecient of the varaibles obtained from all three models A, B and C are comapred to identify that the signs of the coeffiecients are consistent among the three models. Table 2.7 shows that the overall coefficients of model estimated with Datasets A, B and C are consistent and have the same sign in all three models except for winter that is not in any case significantly different from zero. The *T* test was used to compare the coefficients of Datasets B and C, T_{BC} values were estimated by using following formula:

$$T_{BC} = \frac{\theta_B - \theta_C}{\sqrt{S_B{}^2 + S_C{}^2}} \qquad 2\text{-}32$$

where θ_B and θ_C are the estimated coefficients from Dataset B and C and S_B and S_C are the corresponding standarad errors.

It is found from the T_{BC} test values that the coefficients of model B are not significantly different from the coefficients of model C as all the estimated values of T_{BC} are less than 1.96. The coefficients and their *t* values are given in Table 2.7 and are shown graphically in Figure 2.7. It is to note that the presented coefficients are obtained by using the deviation coding as explained in section 2.5.2 due to which the coefficient represents the comparsion with reference to the group mean rather than a particular reference category as in the case of simple coding. We note here that because deviation coding is used here, the coefficients of factors have zero sum. Due to this coding structure, the coefficient of Saturday will be equal to the minus sum of all other days (Weekday and Sunday). Similarly the coefficient of Spring will be equal to the minus sum of all other seasons (Summer, Autumn and Winter). Same procedure is applied to estimate the coefficients of remaining interaction terms.

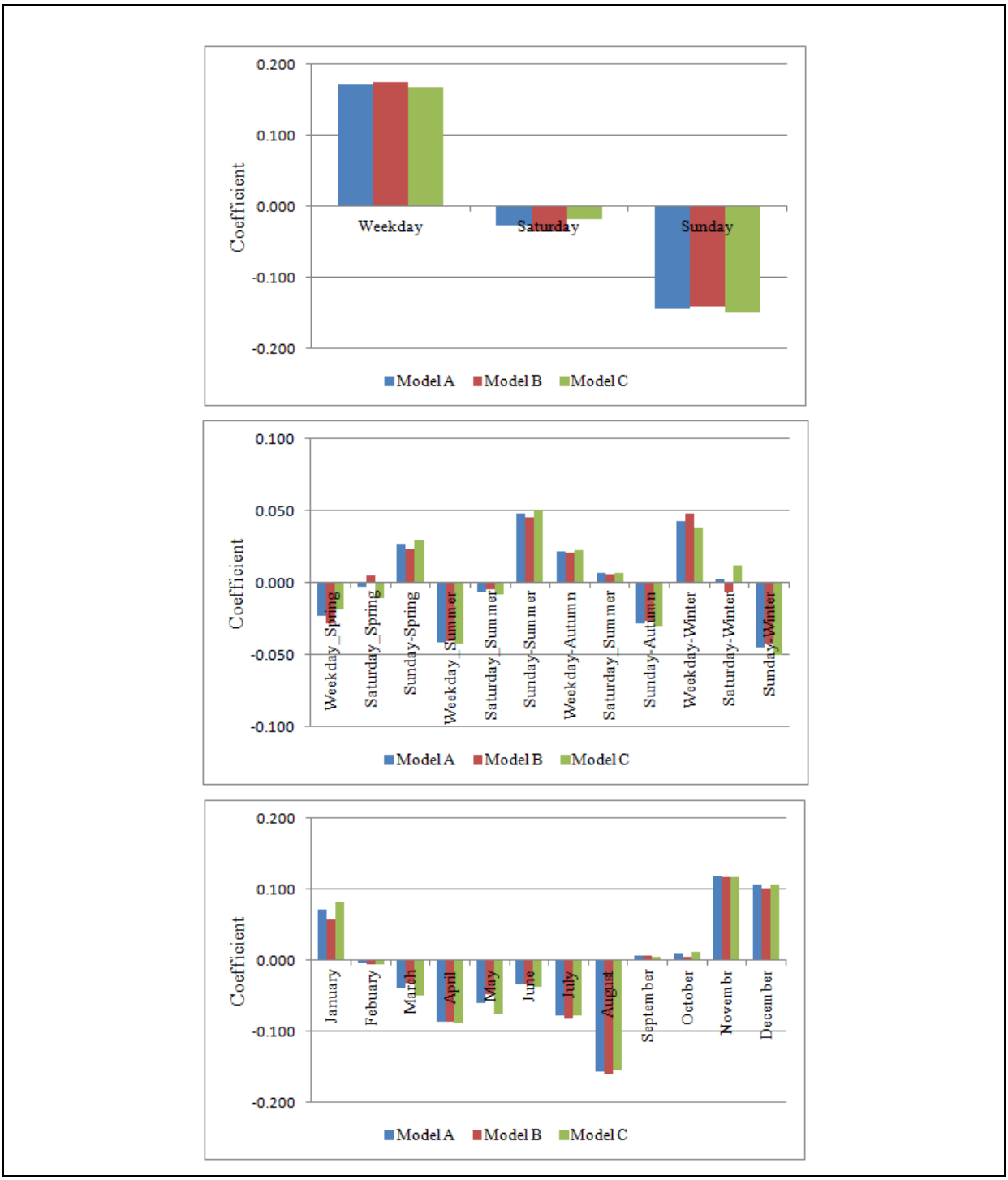
The summary of the comparsion of the estimated coefficients of model A, B and C is as follows:

- The coefficient of the weekday and Sunday variables had significant *t* values and consistent coefficient signs in all three models.
- The coefficient of summer is negative and significant in all three models. Winter is non-significant among the three models while the coefficient of autumn is significant in model B and C only.
- The coefficients of interaction variables of weekday 3 and season were significant and had the same sign in all three models.
- Among the coefficients of month only February and October had non-significant *t* values in model A whereas only May had non-significant *t* values in model C.
- The coefficient of time, Public holidays, Christmas holidays, New-year holidays and distance travelled per vehicle had significant *t* values in all three models.

Table 2.6: Split sample validation results for Dataset 1

Split sample validation				
		Model coefficients *(k=25)*		
Data		A	B	C
A		$\mathbf{x}_A' \boldsymbol{\beta}_A$		
	n	5,479		
	Likelihood	-32,483		
	Deviance	5,502		
B			$\mathbf{x}_B' \boldsymbol{\beta}_B$	$\mathbf{x}_B' \boldsymbol{\beta}_C$
	n		2,739	2,739
	Likelihood		-16,244	-16,265
	Deviance		2,751	2,856
C			$\mathbf{x}_C' \boldsymbol{\beta}_B$	$\mathbf{x}_C' \boldsymbol{\beta}_C$
	n		2,740	2,740
	Likelihood		-16,251	-16,229
	Deviance		2,793	2,751
Total	Likelihood	-32,483	-32,495	-32,494
	Deviance	5,502	5,544	5,607

Figure 2.7: Comparison of coefficient of GLM-Model 19-NB for coefficient validation (Dataset 1)



In graph the coefficients of month represents the combined effect of month and season.

Table 2.7: Comparison of coefficients and *t* values of GLM-Model 19-NB for coefficient validation (Dataset 1)

Variables	Comparison of the coefficients and *t* values of the Models						
	Model A		Model B		Model C		T test
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	T_{BC}
Weekday	0.172	56.13	0.176	40.82	0.168	38.55	1.296
Sunday	-0.145	-34.71	-0.141	-23.99	-0.150	-24.98	1.124
Summer	-0.034	-3.55	-0.034	-2.74	-0.037	-3.02	0.185
Autumn	*0.111*	*1.92*	0.127	5.48	0.130	5.85	-0.107
Winter	*0.009*	*0.19*	*-0.006*	*-0.50*	*-0.005*	*-0.46*	-0.062
Weekday-Summer	-0.041	-8.70	-0.041	-6.09	-0.042	-6.22	0.133
Sunday-Summer	0.048	7.35	0.046	5.08	0.050	5.34	-0.372
Weekday-Autumn	0.022	3.67	0.021	2.48	0.023	2.73	-0.162
Sunday-Autumn	-0.029	-3.53	-0.027	-2.32	-0.030	-2.61	0.195
Weekday-Winter	0.043	8.17	0.048	6.54	0.038	5.07	0.968
Sunday-Winter	-0.045	-6.34	-0.042	-4.23	-0.050	-4.84	0.574
January	*0.063*	*1.46*	0.064	4.36	0.088	6.39	-1.183
February	*-0.013*	*-0.30*	*This variable is dropped from the model in dataset B and C*				
March	0.047	4.76	0.055	4.08	0.039	2.69	0.828
May	0.026	2.64	0.040	2.90	*0.012*	*0.86*	1.393
July	-0.044	-4.50	-0.048	-3.39	-0.041	-3.01	-0.360
August	-0.123	-12.54	-0.128	-9.28	-0.119	-8.57	-0.436
September	0.041	4.14	0.040	2.84	0.041	2.97	-0.050
October	*-0.101*	*-1.59*	-0.122	-4.81	-0.119	-4.83	-0.087
December	0.097	2.28	0.108	7.50	0.112	7.95	-0.185
Time	-3.04E-05	-9.71	-2.9E-05	-6.52	-3.1E-05	-7.11	0.318
Public Holidays	-0.240	-15.59	-0.218	-10.03	-0.262	-11.98	1.437
Christmas Holidays	-0.556	-16.94	-0.565	-12.48	-0.551	-11.56	-0.202
New-year Holidays	-0.223	-5.76	-0.310	-5.28	-0.165	-3.20	-1.854
D.T per veh*	0.00012	13.07	0.00012	9.10	0.00012	9.45	-0.263
Constant	-16.464	-104.34	-16.439	-73.54	-16.507	-73.79	0.215

**D.T per veh= Distance travelled per vehicle,*

Italic shows that these variables are not significant at 5 percent level.

2.6.2.5 Durbin-Watson test

Because the dataset contains cross-sectional time-series data, the possibility arises that serial correlation exists. If this arises, the *t* values of the GLM coefficient would be affected. The Durbin-Watson test was carried out to check whether the autocorrelation exists among the residuals. The presence of autocorrelation was tested in both the whole dataset and for the observations in each of the years. The formula given in equation 2-30 is used to calculate the values of Durbin-Watson statistics. The lower d_l and upper d_u critical values of Durbin-Watson statistics were obtained from the reference values in Table 2.2 by using the number of observations and number of variables in the regression equation. The respective values of d_l and d_u were 1.57 and 1.78. The residuals of model 19 with the generalized linear model using negative binomial gave the estimated value of Durbin-Watson statistics to be 1.03 which lies in the first region between 0 and 1.57. This identifies the presence of positive autocorrelation in the data so that the null hypothesis for the absence of autocorrelation was rejected. The Durbin-Watson statistic was also calculated for each year. Based on the test results, the null hypothesis for the absence of autocorrelation among residuals was rejected for each of the 15 years. The results given in Table 2.8 show that residuals are autocorrelated and serial correlation exists within each year.

Table 2.8: Durbin-Watson test results for Dataset 1

Observation	Year	DW	Observation	Year	DW
1	1991	1.02	9	1999	0.64
2	1992	1.24	10	2000	0.93
3	1993	0.98	11	2001	0.81
4	1994	1.06	12	2002	1.23
5	1995	0.99	13	2003	1.18
6	1996	1.23	14	2004	0.97
7	1997	1.08	15	2005	1.07
8	1998	1.03			

DW represents the Durbin Watson statistic

2.6.2.6 Preferred model

Model 19 was preferred based on the results obtained in section 2.6.2.1-5. In the model selection process, *BIC* values were compared which were used to guide rather than to dictate the selection of a model along with other considerations for model selection. The detail of model preference was based on the criteria set in section 2.5.4 for assessment of model performance.

From section 2.6.2.1.2, model 19 was identified as having good *BIC*, significant coefficients of most of explanatory variables including time and circumstantial variable of distance travelled per vehicle. Tests for multicollinearity of the explanatory variables in this model using the VIF showed these variables have acceptable value. Other models (model 17 and 18) in which circumstantial variables of vehicle per person and distance travelled per person respectively were investigated did not have better *BIC* values and multicollinearity existed between the time and these circumstantial variables. It was also observed that when the circumstantial variables were used together in different combinations in model 20-26, it resulted in improvement of *BIC* in some cases, but time and circumstantial variables had high VIF, as a result those models were not preferred. Analysis of temporal effects in section 2.6.2.2 also showed that in model 19 no substantial systematic temporal trend remains that can be represented by further quadratic temporal terms in the model. Due to this, model 19 was carried forward for the split sample analysis to validate and check the consistency of the model and its parameter estimates.

Split-sample tests reported in Section 2.6.2.4 showed the estimates of parameters for model 19 to be consistent and reliable. After this, the Durbin-Watson test was used to test for the presence of serial correlation in the residuals of the model 19. However, it was found in section 2.6.2.5 that serial correlation exists in the residuals of this model (Table 2.8). Due to this, the Generalized Estimation Equation (GEE) with autoregressive (AR1) error term for model 19 was therefore preferred over the GLM because it can accommodate this serial correlation.

In section 2.6.2.6.1 the coefficients of Model 19 with GEE-AR1-negative binomial are compared with GLM-negative binomial to identify the extent to which estimates and significance level of the coefficients differ among these model forms. Further analysis was

carried out in the coming sections on the results obtained from the preferred model (Model 19 with generalized estimation equation with negative binomial and having AR1 error structure).

2.6.2.6.1 Comparison of coefficients for Dataset 1 (GEE and GLM)

Stata software was used to estimate the coefficients of all variables which were found to have expected signs. The comparison of the coefficients and *t* values for model 19 by using GEE with negative binomial having autoregressive error structure (AR1) and GLM with negative binomial was carried out. Because both models were fitted to the same data, the estimates of corresponding parameters are not mutually independent. Because of this, no formal *T* test could be undertaken between the values estimated by the different models, so an informal comparison is presented here instead. In all cases the coefficient and their sign remained same in both the models. However, a slight change in the *t* values was observed. It was found that the *t* values of the variables of weekday, Sunday, all the interaction variables and Public holidays have increased in GEE while for the month, time, Christmas holidays, New-year holidays and distance travelled per vehicle variable their *t* values have decreased. This might be due to presence of serial correlation in the data.

In this model the distance travelled profiled by each day which takes into account the variations by day of week and month of year is used in the offset. Due to this, the coefficients of weekday 3 and month will directly represent their influence on the risk per unit of travel. However, no correction factors for the Public holidays, Christmas holidays and New-year holidays were available. Because of this, the coefficients of these variables represent their influences on the frequency of road accidents rather than risk per unit of travel.

From the estimated coefficients, strong effects on the risk per unit of travel were identified for weekday, Sunday, interaction between seasons and weekday 3, time, month and distance travelled per vehicle. It was found that the coefficients of autumn, winter, January, February and October had non-significant *t* values in both the models whereas summer, May and December which had significant *t* values in GLM turned to be non-significant in GEE model. Generally it is observed that some coefficients may differ in value in GEE and GLM and the accuracy of the estimation also differ. The coefficients estimated by GEE-AR1 which are shown in Table 2.9 are preferred as it can accommodate the presence of serial correlation.

From model results it was observed that weekday had greatest risk per unit of travel, Saturday had about 20 percent lower risk and Sunday about 35 percent lower risk than weekdays. The combined effect of month and season showed that November had the greatest risk per unit of travel (about 11 percent greater) than average whereas August had least risk per unit of travel (about 15 percent lower) than the average. The interaction variable of weekday 3 and seasons ranged from 0.048 (Sunday-Summer) to -0.050 (Sunday-Winter). These represent respectively an increase and decrease of about 5 percent in risk per unit of distance travel. The variables of Public holidays, Christmas holidays and New-year holidays had a coefficient of -0.216, -0.426 and -0.116 respectively which represents variation in frequencies of road accident occurrence on these days rather than risk. The coefficient of time is $-3x10^{-5}$ per day, which shows that the risk per unit of distance travelled had decreased at about 1 percent per annum. The distance travelled per vehicle variable has a positive coefficient which shows that the years in which fewer vehicles were registered there was a greater risk of road accident involvement per unit of distance travelled.

After this the estimated coefficients of weekday 3, seasons, interaction of weekday 3 and seasons, and month were combined together to give an understanding of the combined effect. Figure 2.8 shows the comparison of the risk per unit of distance travel on weekday, Saturday and Sunday by month of year. It is observed that risk per unit of travel was greater for autumn and winter months. Weekdays had greater risk than Saturday and Sunday when compared within each month. For weekdays the risk per unit of distance travel is greater in the months of November to January (about 20 percent greater than in months of April-July). During the summer month it fluctuates but in September it increases sharply. Sunday had the lowest risk per unit of travel of all the days of week: it has least risk in August which is about 18 percent lower than Sunday in November. Further associations are also observed which show that Saturday in winter has greater risk per unit of travel than some of the weekdays in spring and summer. Saturdays in November had slightly greater risk per unit of travel than weekdays in April and July.

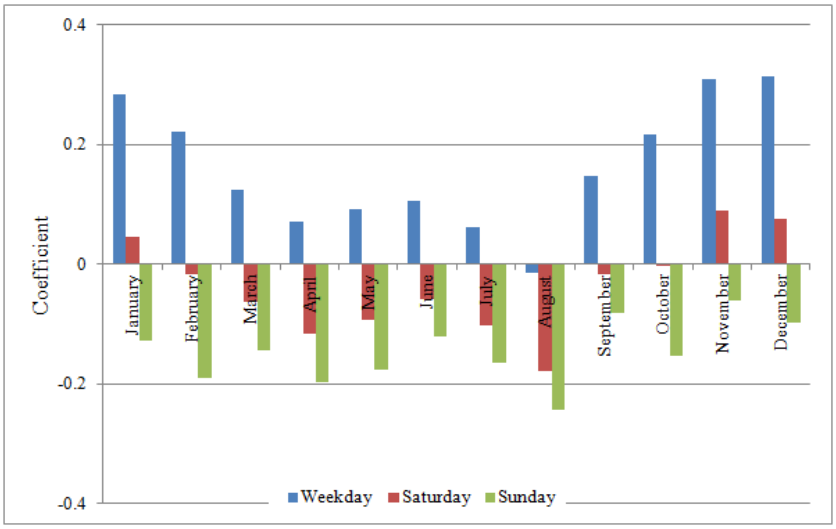
Table 2.9: Comparison of coefficients and *t* values of model 19 (GEE-AR1 and GLM) negative binomial for coefficient validation (Dataset 1)

Variables	Comparison of the coefficients and *t* values of the Models			
	GEE-AR1-NB		GLM-NB	
	Coefficient	*t* value	Coefficient	*t* value
Weekday	0.174	63.58	0.172	56.13
Sunday	-0.146	-50.01	-0.145	-34.71
Summer	*-0.024*	*-1.84*	-0.034	-3.55
Autumn	*0.068*	*1.47*	*0.111*	*1.92*
Winter	*0.039*	*1.01*	*0.009*	*0.19*
Weekday-Summer	-0.043	-10.27	-0.041	-8.70
Sunday-Summer	0.048	10.71	0.048	7.35
Weekday-Autumn	0.023	4.37	0.022	3.67
Sunday-Autumn	-0.028	-4.87	-0.029	-3.53
Weekday-Winter	0.043	9.22	0.043	8.17
Sunday-Winter	-0.050	-9.93	-0.045	-6.34
January	*0.028*	*0.77*	*0.063*	*1.46*
February	*-0.034*	*-0.96*	*-0.013*	*-0.30*
March	0.054	3.47	0.047	4.76
May	*0.023*	*1.47*	0.026	2.64
July	-0.044	-2.87	-0.044	-4.50
August	-0.122	-7.68	-0.123	-12.54
September	0.041	2.55	0.041	4.14
October	*-0.047*	*-0.92*	*-0.101*	*-1.59*
December	*0.058*	*1.66*	0.097	2.28
Time	-3.09E-05	-5.75	-3.04E-05	-9.71
Public Holidays	-0.216	-16.85	-0.240	-15.59
Christmas Holidays	-0.426	-14.03	-0.556	-16.94
New-year Holidays	-0.116	-3.51	-0.223	-5.76
D.T per veh*	0.00012	7.60	0.00012	13.07
Constant	-16.461	-61.01	-16.464	-104.34

* Distance travelled per vehicle,

Italic shows that these variables are not significant at 5 percent level.

Figure 2.8: Comparison of risk per unit of distance travelled on Weekday, Saturday and Sunday by month of year (Dataset 1)



2.6.2.6.2 Comparison of number of road accidents observed and estimated, Standardized deviance residuals and cumulative percentage graphs:

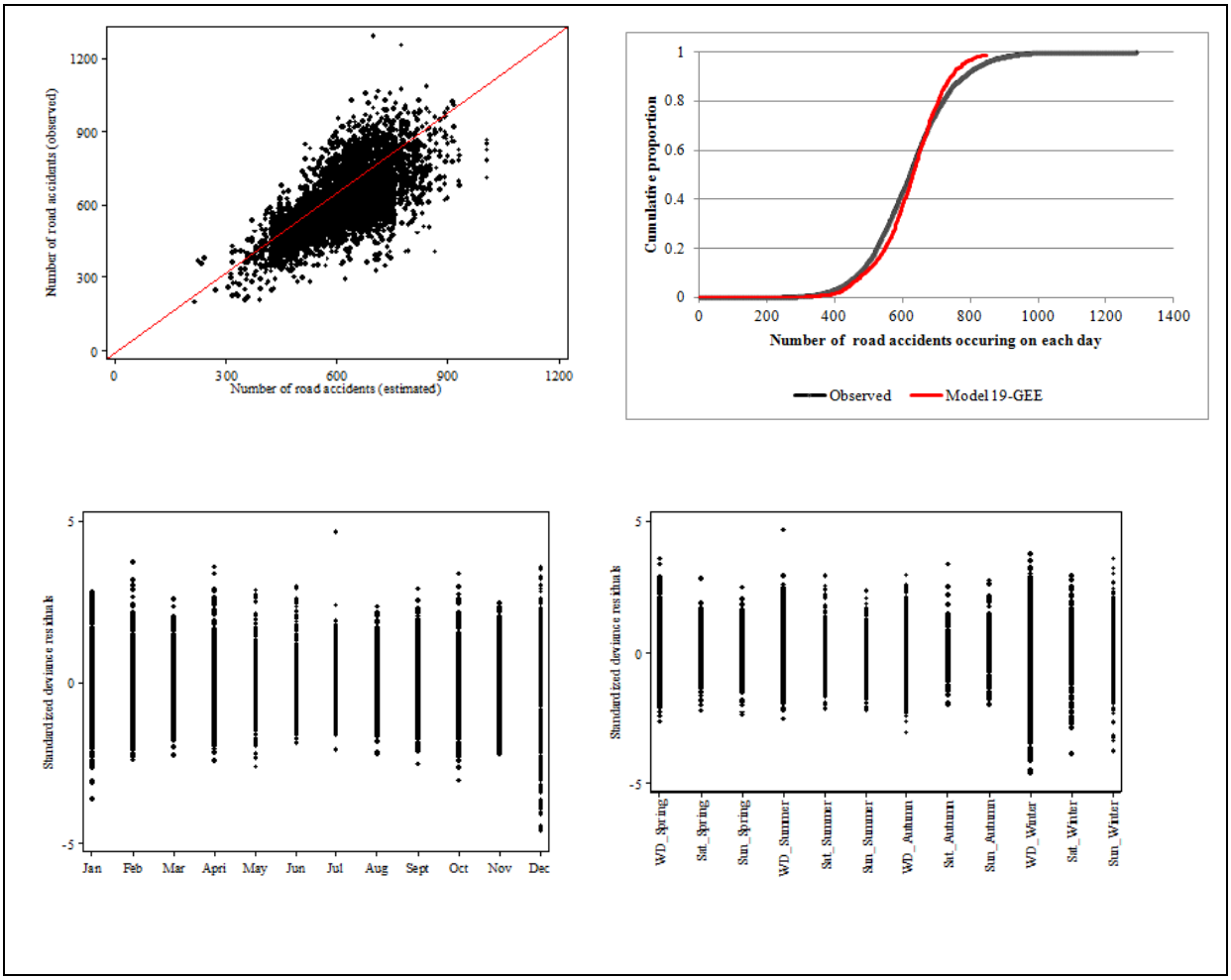
Graphs of observed values of road accidents against the estimated value by model 19 with GEE negative binomial having AR1 error structure for Dataset 1 (whole of Great Britain) are presented in Figure 2.9. The graph of road accidents observed and estimated shows that model have generally represented the data well as the line of equality passes through the centre. However, the cumulative proportion graph shows that the model estimated slightly fewer observations with number of road accidents less than 600 and a greater number of observations with number of accidents greater than 600 in comparison to the observed data.

From the graphs and further exploration of data it was found that the two days with the highest SDRs were 3rd July 1992 and 1st February 1991, both Fridays (weekdays) which gave the standardized residual deviance of 4.65 and 3.74 respectively. The number of road accidents observed on these days was 1,290 and 848 whereas the estimated values for these days were 699 and 511 respectively.

The standardized deviance residual (SDR) graph showed that the SDR generally remained between +4 and -4. The graphs of standardized deviance residuals plotted against month showed December and January had highest range of SDR among months, even after including Public holidays, Christmas, and New-Year holiday variables in the model. In the

same way weekdays, Saturday and Sunday in winter had higher SDR than all other combinations of weekday and seasons. Upon investigation it was found that among the highest hundred negative SDRs, 60 observations belonged to the December and 17 belonged to January mostly relating to the dates between the Christmas and New-year holidays. This suggests that the model is not able to precisely estimate the number of road accidents for that period. It was also observed that the range of SDR of the months of July and August is smaller than other months which reveal that this model can estimate the number of road accidents for these months more accurately than other months so that these are less variable than other months and thus easier to model as a whole.

Figure 2.9: Number of road accidents observed and estimated, Cumulative proportion and Standardized deviance residuals graphs (Dataset 1)



2.6.2.6.3 Final model checking graphs

For model checking the following four diagnostic plots were produced, as shown in Figure 2.10.

1. Plot of deviance residuals against fitted values
2. Normal quantile plot
3. Scale location plot
4. Cook's distance plot

In the first graph the deviance residuals produced by model 19 with GEE-AR1 negative binomial are plotted against fitted values: this does not show any trend. Attention was paid to identify any increase in the deviance with increase in the fitted values, because higher fitted values that had higher deviance would have been a cause of concern. This graph shows that the model is correct as deviance is scattered evenly around the zero line.

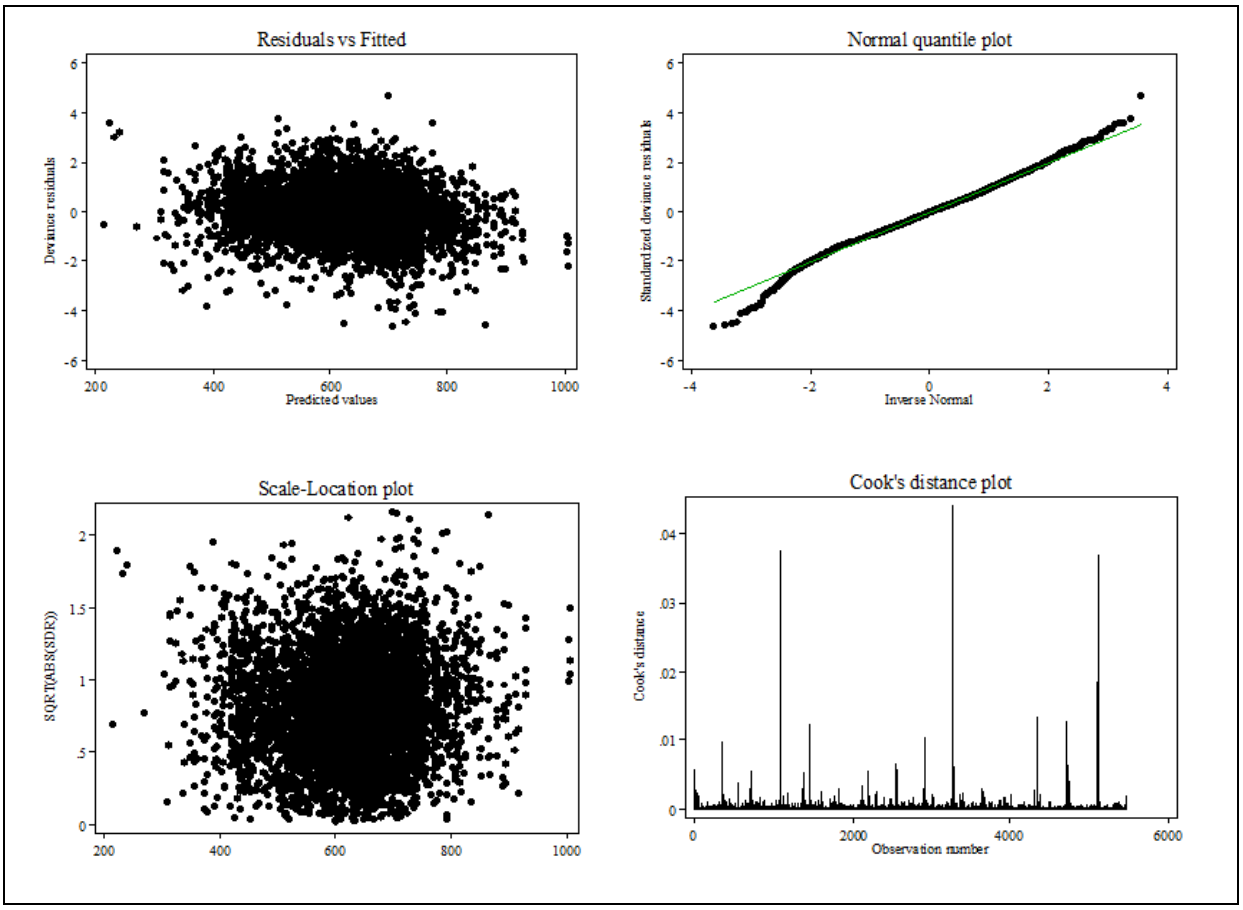
In the second graph the normal quantile plot of standardized deviance residuals is shown. This was used as a diagnostic tool to check that the deviance residuals have a distribution close to normal. The graph shows that for much of the range the quantile plot follows a reference line which verifies the assumptions of normality of the residuals. However, a low deviation is observed at the low end of the range, which suggests that the data distribution has relatively few observations that fit closely. In the third graph, the scale location plot which is the repeat of first but on a different scale, it shows the square root of the absolute value of SDR against the fitted value. This shows that variance does not increase with the increase in mean.

In the last graph, Cook's distance shows the observations which have the most influence upon the fitted model and if these observations were excluded, the parameter estimates will change a lot. In order to find out the higher peaks the dataset was investigated, it was found out that most of the observations that had higher peaks corresponded to December and January values (25^{th}, 26^{th} December, 1^{st} January). A critical value of 1 (Montgomery, 2010) was considered to be a cut off value for Cook's distance which would indicate that the observation is influential and its removal will result in changing the coefficient value considerably. However, in this case the values of observations are in the range (all are less

than 0.05) that does not cause problems. Due to this, no observation was removed from the dataset.

In order to verify the assumption of homoscedasticity (equal variance) in the residuals of the model 19 GEE-AR1, two different tests (Park Test and Glejser Test) as suggested by Gujarati (2009) were used to check the presence of heteroscedasticity in the residuals. The details about the procedure of these tests are given in Gujarati (2009, page 396). The results of these two tests showed that the *t* values of the estimated number of road accidents was found to be non-significant when regressed against the squared values of residuals for Park test and absolute values of residuals by Glejser Test. These results suggest that heteroscedasticity is not present in the residuals of this model which verified the assumption of homoscedasticity. The results of the tests are shown in Appendix A2.6.

Figure 2.10: Diagnostic plots for model 19 (Dataset 1)



2.6.3 Model selection process, goodness of fit and model checks for Dataset 2

Dataset 1 was found to be over-dispersed in the sense that the variances of the residuals exceeded the estimated value, even with respect to the most detailed models. Due to this, the negative binomial error structure was preferred to Poisson for regression. However, in order to further explore the differences in the number of road accidents in various geographical areas another approach was made by disaggregating the dataset to police force level so that further information that is available at this level from the Office for National Statistics, Department for Transport and from the STATS 19 form could be incorporated into the model. Disaggregating the national data in this way is likely to lead to correlation between similarly located observations made at the same time due to common regional effects such as weather. Due to the effect of spatial autocorrelation, the data from different police forces at the same time can not be regarded as mutually independent. In the present study these police forces were treated as independent and no adjustment was made for this. However, it is understood that this could lead to underestimation of standard errors of model parameters (by factor typically in the range of 1.5-2.5), and hence overestimation of their associated *t* values. Due to this, when identifying the effect of an explanatory variable as significant, its *t* value was considered with caution.

There are two main possibilities for area-specific disaggregation levels of STATS 19 data for the whole of Great Britain which are either by police force or by local authority. Dataset 1 was disaggregated to police force level, with 51 values. This was preferred to disaggregation to the finer level of local authority level, which would have generated a very large dataset.

A new dataset was created which consisted of number of road accidents on each day recorded by each police force from 1st January 1991 to 31st December 2005. Each police force represented a single or group of local authorities. This increased the number of observations to 279,429 in Dataset 2 in comparison to 5,479 observations in Dataset 1. The information about population, length of all roads, length of all classes of road, population density and number of registered vehicles was obtained for each local council from the Office for National Statistics and the Department for Transport. This data was then aggregated to police force level by using STATS 20 ‘Instructions for the completion of road accident reports’ which showed all the local councils in the particular police force area. From this information circumstantial variables such as vehicles per person, vehicles per road length, vehicles per

surface area and length of each road class as a proportion of the total road length in a police force area were derived. As with Dataset 1 it was found that negative binomial performed better than Poisson regression, indicating that over-dispersion remains, and due to this only negative binomial was used for model development for Dataset 2.

2.6.3.1 Negative binomial regression model for 51 police force areas of Great Britain (Dataset 2)

2.6.3.1.1 Negative binomial regression model

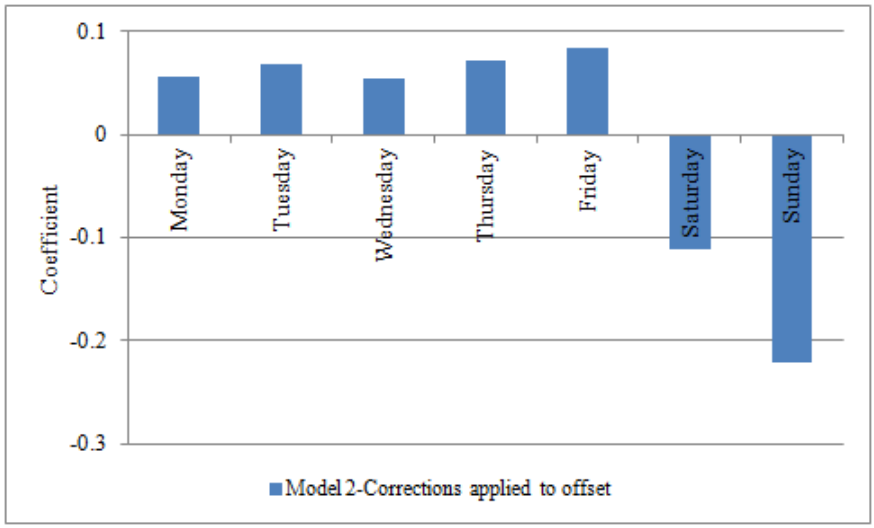
For Dataset 2, a total of 33 models were developed with different combinations of variables as shown in Figure 2.12. The initial model was developed with only a constant term. The incremental procedure was applied for estimating the mean of the number of road accidents for each day by police force in Great Britain. An offset variable was also used in each of these models. As described in section 2.5.3 two different variables were considered for use as the offset variable. Initially the models were developed with the logarithm of the national vehicle-kilometres of road travel as an offset. This variable does not distinguish among the police force areas but does allow for different usage for the day of week, month and years. The log-likelihood and *BIC* values of these models are shown in appendix Table A2.2.

It is understood that the distance travelled on each day varies by day of week, month and police force. As the unit of observation in dataset 2 is number of road accidents on each day for each of the police force, due to this an adjustment was made in national vehicle kilometres to account for the variations in distance travelled among the police forces. As a result, a new variable was derived by using the information of number of registered vehicles in each police force area, total national number of vehicles and national vehicle kilometres. It was assumed that vehicle kilometres travelled within each police force area is proportional to the number of registered vehicles there. The details of this are given in section 2.5.3. After this, correction factors for day of week and month were used for the offset variable to account for the variation in distance travelled. The use of the profiled distance travelled in offset results in direct interpretation of the estimated coefficients as risk per unit of travel. This variable when used in offset produced better *BIC* results than the national vehicle kilometres which was unable to account variations in distance travelled among police forces. Due to this, it was preferred to be used as an offset.

It is to be noted that in the process of model development the corrections to adjust the distance travelled by day of week and month were applied to the offset only when the related variables were introduced into the model as explanatory variable. Table 2.3 given in section 2.6.2.1.1 shows the list of models and the corrections applied to offset.

The initial model gave *BIC* value of 1,654,569. It was observed from the results of model 2 when the offset variable was adjusted to take account of the variations in distance travelled by day of week, the estimated risk per unit of distance travel for each of the weekday was found to be quite similar. The comparison of the estimated coefficients from model 2 is shown in Figure 2.11. Keeping these results in view a new variable of weekday 3 was introduced in model 4 with 3 levels each representing weekday, Saturday, and Sunday. In the same way Season (Spring, Summer, Autumn and Winter) were introduced in model 5. Model 4 and 5 are considered to be simple versions of model 2 and 3 respectively.

Figure 2.11: Comparison of the coefficients of day of week from model 2 (Dataset 2)



Comparison of the results showed that model 2 with day of week had better *BIC* (by 56) values than model 4 which had weekday 3 as explanatory variable. In the same way, model 3 with month variable had performed better than model 5 with season. In model 6, day of week and month while in model 7, weekday 3 and season variables were used together. This showed that model 6 had performed better than model 7 in terms of *BIC* values. The *BIC* of model 6 was better by value of 2,349 than model 7.

In model 8 and 9 interaction terms were introduced. Model 8 had *BIC* values of 1,642,540 with 84 degrees of freedom. The *BIC* of model 8 was better by 1,937 than model 9 which was the simple version of model 8 with only 12 degrees of freedom. As it is also evident from the results of model 3 that month variable is also important and we understand that number of road accidents vary by month because of this it was introduced in model 10 along with weekday 3, season and interaction of weekday 3.season variables. Introduction of month variable improved the *BIC* of model 10 in comparison to model 8. Despite having 62 fewer degrees of freedom than model 8 the *BIC* of model 10 was better by 352.

It was observed that the *BIC* of model 2 with day of week and model 4 with weekday 3 variable were better than the model 10. This was because day of week and month corrections to the offset were applied together (in model 10) when these variables were introduced as explanatory variables which results in loss in the value of *BIC*. Based on our understanding that month, Season, their interaction, and monthly adjustments to the offset variable are important and necessary hence these were included into the model. Due to these reasons, model 10 was carried forward instead of model 2 and 4 despite having slightly less preferable *BIC* values.

In model 11 after including the time variable substantial improvement of 13,934 in *BIC* was achieved. The *BIC* of model 14 after introducing the Public holidays, Christmas holidays and New-year holidays was found to be 1,623,944 which have an improvement of about 4,310 in the *BIC* in comparison to model 11. For model 15, a police force specific factor was introduced that subsumes the explanatory function of all area-specific units. For model 15 significant improvements of about 88,000 in the *BIC* value was observed in comparison to model 14.

Due to the disaggregated nature of the data it was possible to introduce more explanatory variables into the model. From model 15 onwards police force specific variables (circumstantial variables) were used to account for all differences among the police forces.

In models 16 to 21 the police force specific variables (circumstantial variables) were introduced individually into the model, out of which model 17 with vehicles per head of population had better *BIC* value. Model 19 and 21 with the variables of vehicles per surface area and ratio of each road class to total road length respectively had also good results. After

this, circumstantial variables were incorporated into the models in various combinations. From models 22 to 26, model 22 with population density and vehicles per head of population produced better *BIC* values. From model 27 to 29, model 29 with ratio of each road class to total road length, population density and vehicle per head of population had better *BIC* values. In model 32 when all the area-specific variables were incorporated into the model, it was observed that *BIC* values were better than model 15 where police variable was used. In model 33, the police force variable was introduced along with all area-specific variables which had better *BIC* values among all models because it has police force variable as a factor, but it lacks explanatory power. Model 33 produced a better fit than model 15 according to the *BIC* because of the temporal variation in the area-specific circumstantial variables. Detailed results for the all models are shown in Table 2.10.

Figure 2.12: Lattice of model development Dataset 2

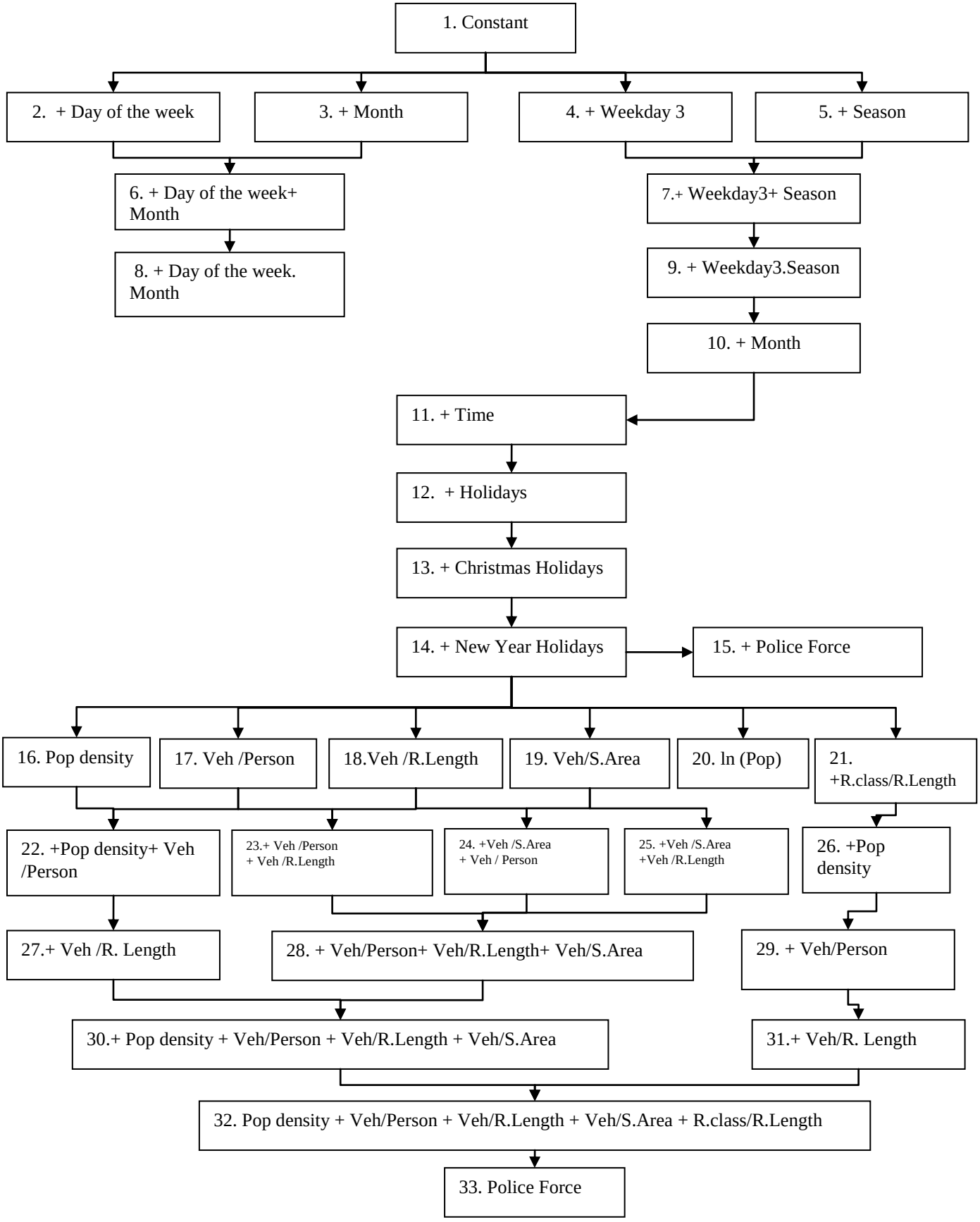


Table 2.10: Results of all models for the 51 police forces of Great Britain (Dataset 2)

Model	**D.F**	**Scale**	**Likelihood**	***BIC***
1	1	0.14777	-827,278	1,654,569
2	7	0.13548	-820,731	1,641,549
3	12	0.14915	-827,913	1,655,976
4	3	0.13555	-820,784	1,641,605
5	4	0.15150	-829,078	1,658,206
6	18	0.13696	-821,435	1,643,095
7	6	0.13933	-822,684	1,645,444
8	84	0.13585	-820,743	1,642,540
9	12	0.13852	-822,163	1,644,477
10	22	0.13621	-820,956	1,642,188
11	23	0.12156	-813,983	1,628,254
12	24	0.11895	-812,511	1,625,323
13	25	0.11810	-811,919	1,624,151
14	26	0.11796	-811,809	1,623,944
15	76	0.06948	-767,349	1,535,651
16	27	0.09529	-800,720	1,601,779
17	27	0.07029	-777,246	1,554,830
18	27	0.11423	-811,050	1,622,439
19	27	0.10238	-805,021	1,610,381
20	27	0.10068	-806,271	1,612,880
21	42	0.08495	-785,760	1,572,047
22	28	0.06195	-772,462	1,545,275
23	28	0.06292	-773,535	1,547,422
24	28	0.06333	-773,251	1,546,854
25	28	0.09504	-799,652	1,599,656
26	43	0.08458	-785,422	1,571,383
27	29	0.06190	-772,444	1,545,251
28	29	0.06289	-773,103	1,546,570
29	44	0.05725	-758,700	1,517,952
30	30	0.05953	-771,227	1,542,830
31	45	0.05657	-758,403	1,517,370
32	46	0.05653	-758,268	1,517,113
33	96	0.04487	-744,848	1,490,900

D.F= degrees of freedom; BIC= Bayesian information criterion

2.6.3.2 Analysing the temporal effects

The procedure presented in section 2.6.2.2 was used to investigate for the presence of further temporal effects that were not represented in the models. For this, time and square of time variables were added to model 1-10 whereas from model 11 onwards only the square of time was added as these models already included a time variable. The resulting improvement in *BIC*, coefficients and *t* values of time and square of time, and their variance inflation factors were examined.

It is observed from the results which are shown in appendix Table A2.7 that huge improvements in the value of *BIC* ranging from about 2,000 to 13,000 were achieved when temporal trend was added to each of the models 1-10, which indicates that these models did not account for temporal effects. In each case the variables of time and square of time variables had significant *t* values but the estimated variance inflation factors were found to be high (value of 16) which suggests that the true effects of time and square of time cannot be identified through their estimated coefficients and standard errors because of multicollinearity.

From model 10 onwards the improvement in *BIC* on inclusion of the square of time was smaller (in range of 26 and 179) which is because these models already include a time variable: this suggest that these models already include most of the temporal effects by the use of time and other explanatory variables. Model 33 with 96 degrees of freedom, which had the better *BIC* value than other models, showed no improvement in *BIC* after adding square of time (one degree of freedom), though an improvement of 3 was observed in the value of log-likelihood which shows that temporal trend has already been represented adequately by the model.

These tests show that models 1-10 do not have an adequate representation of time. Model 11-33 have a good representation through the linear time variable and other explanatory variables which vary over time, only a small improvement in model performance can be achieved by allowing for further variation over time according to a quadratic term.

2.6.3.3 Checking for the presence of multicollinearity

Variance inflation factors were estimated for the models in order to investigate the presence of multicollinearity among the explanatory variables. It is expected that due to the nature of the data and associations among the explanatory variables some of them will necessarily have high VIFs. On the other hand where multicollinearity exists among the variables of time, population density, vehicles per person, vehicle per road length, vehicle per surface area, population and proportion of road length by road class than it will be difficult to identify the true effects of each of these variables individually. Models in which high VIFs (greater than 10) were estimated for these variables were not preferred because this shows that the associated variables added relatively little information. The results in Table 2.11 are presented as values of VIFs for some of the circumstantial variables which are used in model 16-33.

Table 2.11 shows that circumstantial variables such as population density, vehicles per head of population, vehicles per kilometre of road length and vehicles per square kilometre of surface area when used individually in model 16-19 produced low VIF values showing that these variables do not have strong temporal trends. In model 21 the variable of ratio of road class to road length (with 16 degrees of freedom) had high VIF.

From model 22 onwards these area-specific and other variables were used jointly. Models 26, 29, 31, 32 and 33 included the variable of ratio of road class to road length that had high VIF (greater than 40). Models 22-25, 27 and 28 had acceptable values of VIF for individual variables. In model 22 population density and vehicles per head of population have low VIF of 1.1 and 1.4 respectively. In model 27 where population density, vehicles per person and vehicles per road length were used together, these variables had low VIF of 2.1, 1.8 and 2.2 respectively. From model 29 onwards, where area-specific circumstantial variables were used together in different combinations, unacceptable high VIF values were observed which indicate that the joint use of these variables will result in multicollinearity.

It was observed from the table 2.11 that the models 22-25, 27 and 28 have acceptable values of VIF for the time, population density, vehicles per person, vehicles per road length and vehicles per surface area. Among these model 22 and 27 had better *BIC* values. Model 27 was not considered further for the reasons that are explained in section 2.6.3.6. From these,

model 22 was carried forward for split sample analysis as it has good *BIC* and acceptable VIF values.

Table 2.11: Variance inflation factors VIF of variables for Dataset 2

Model	Time	P.D	V/P	V/R	V/A	Ln(P)	*Mean R.C/R
16	1.0	1.0					
17	1.3		1.3				
18	1.1			1.1			
19	1.0				1.0		
20	1.0					1.0	
21	1.2						42.8
22	1.3	1.1	1.4				
23	1.3		1.4	1.1			
24	1.3		1.3		1.0		
25	1.1			2.0	1.9		
26	1.2	4.6					43.8
27	1.3	2.1	1.8	2.2			
28	1.3		1.6	2.4	2.3		
29	1.7	2.15	1.8				44.1
30	1.3	38.8	2.2	2.5	41.2		
31	1.7	2.5	5.1	15.3			45.1
32	1.7	54.5	5.2	17.7	56.9		46.0
33	3.7	984.1	18.7	124.3	228.0		90362

P.D=Population density, V/P= vehicles per head of population, V/R= Vehicles per kilometre of road length, V/A=Vehicles per square kilometre of surface area, P= Population, R.C/R= Length of road by class by total road length

2.6.3.4 Split sample tests

After analysing the *BIC*, temporal effects and *VIF* values according to the criteria discussed in section 2.5.4, model 22 was taken forward for further investigation. In order to check the consistency of the model and its parameters, split sample validation tests were undertaken following the same procedure detailed in section 2.6.2.4. The following datasets were used to cross-check and validate the results of model 22.

Full dataset = Data A
Dataset first portion = Data B
Dataset second portion = Data C

The results in Table 2.12 show that values of log-likelihood and total deviance are consistent and do not differ widely between Datasets B and C. The maximised log-likelihood for Datasets B and C was -385,634 and -386,813 respectively. After this the coefficients that were fitted to Datasets B and C were evaluated using log-likelihood and deviance values achieved using the complementary part of the dataset. This produced log-likelihood and deviance values that were slightly worse than those achieved using the data-specific coefficients. The coefficients of Dataset C when used with Dataset B produced the likelihood of -385,653 which differed only by 19 from the value optimised for that dataset. Because the model parameters are not optimised in this case, there are 28 more degrees of freedom in the residuals: this gives a likelihood ratio test statistic of 38 on 28 degrees of freedom, which is less than the critical value of 41.34 at 0.05 significance level. Therefore the null hypothesis cannot be rejected that parameters fitted to Dataset C are as appropriate for Dataset B as these fitted to that dataset. In the same way when coefficients of Dataset B were used with Dataset C it produced a difference of 20 in the likelihood value: this gives likelihood ratio statistic of 40 which is less than critical value of 41.34 at 0.05 level. Table 2.12 shows that the log-likelihood and total deviance values of Dataset A are only marginally better than the sum of the two corresponding values. This confirms that the parameters of model 22 are consistent.

Table 2.12: Split sample validation results for Dataset 2

Split sample validation				
		Model coefficients (k =28)		
Data		A	B	C
A		$\mathbf{x}'_A \boldsymbol{\beta}_A$		
	n	279,429		
	Likelihood	-772,462		
	Deviance	319,880		
B			$\mathbf{x}'_B \boldsymbol{\beta}_B$	$\mathbf{x}'_B \boldsymbol{\beta}_C$
	n		139,715	139,715
	Likelihood		-385,634	-385,653
	Deviance		159,473	159,530
C			$\mathbf{x}'_C \boldsymbol{\beta}_B$	$\mathbf{x}'_C \boldsymbol{\beta}_C$
	n		139,714	139,714
	Likelihood		-386,833	-386,813
	Deviance		160,461	160,403
Total	Likelihood	-772,462	-772,467	-772,466
	Deviance	319,880	319,934	319,933

In the second step the coefficients of Datasets A, B and C are compared which indicates that overall the coefficients of dataset A, B and C are consistent and have the same sign and similar values in all three models. The *T* test was used to compare the coefficients of Datasets B and C because they are fitted to distinct datasets, they are mutually independent. T_{BC} values were estimated by using the formula 2-32. It is found from *T* test values that coefficients of model B are not significantly different from the coefficients of model C as the estimated values of T_{BC} are less than 1.96 except one interaction variable of Sunday-Summer. The comparsion of coefficients and *t* values are shown in Figure 2.13 and Table 2.13. The summary of comparsion is shown below.

- The coefficient of the Weekday and Sunday had significant *t* values and expected signs in all three models.

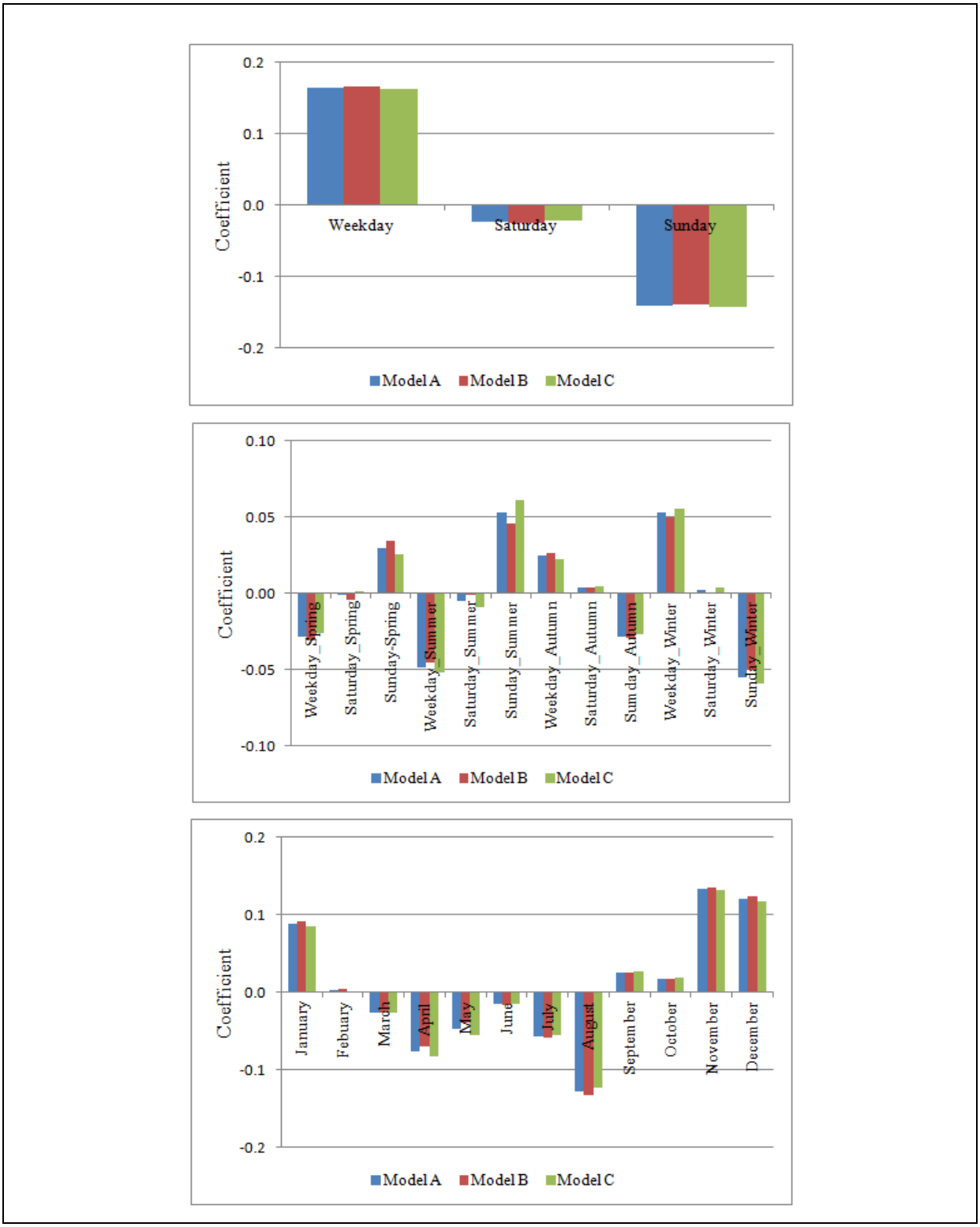
- Among the coefficients of season only Summer and Autumn have significant *t* value in model A.

- All the interaction variables of weekday 3 and season have significant *t* values in all three models.

- Among the coefficients of month Febuary, September and October had non-significant *t* values in all three models.

- The coefficient of Time, Public holidays, Christmas holidays, New-Year holidays, population denisty and vehicles per head of population had similar signs and had significant *t v*alues in all three models.

Figure 2.13: Comparison of coefficient of GLM-Model 22-NB for coefficient validation



In graph the coefficients of month represents the combined effect of month and season.

Table 2.13: Comparison of coefficient and *t* values of GLM-Model 22-NB for coefficient validation

Variables	Comparison of the coefficients and *t* values of the Models						
	Model A		Model B		Model C		T test
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	T_{BC}
Weekday	0.164	135.54	0.166	96.85	0.163	94.86	0.971
Sunday	-0.141	-82.33	-0.140	-58.00	-0.142	-58.44	0.713
Summer	0.085	2.20	*0.085*	*1.53*	*0.085*	*1.59*	-0.003
Autumn	0.049	2.00	0.034	0.97	*0.064*	*1.87*	-0.606
Winter	*-0.057*	*-1.67*	*-0.049*	*-1.01*	*-0.065*	*-1.37*	0.239
Weekday-Summer	-0.049	-26.01	-0.046	-17.29	-0.052	-19.47	1.583
Sunday-Summer	0.053	20.33	0.046	12.33	0.061	16.38	-2.874
Weekday-Autumn	0.024	10.48	0.026	8.00	0.023	6.79	0.816
Sunday-Autumn	-0.028	-8.57	-0.030	-6.39	-0.027	-5.72	-0.443
Weekday-Winter	0.053	25.13	0.050	16.92	0.055	18.63	-1.309
Sunday-Winter	-0.055	-18.45	-0.051	-12.11	-0.059	-13.97	1.476
January	0.145	3.91	0.140	2.64	0.151	2.90	-0.148
February	*0.060*	*1.61*	*0.053*	*1.00*	*0.067*	*1.29*	-0.185
March	0.050	12.88	0.043	7.78	0.057	10.38	-1.825
May	0.029	7.52	0.030	5.47	0.028	5.15	0.227
June	-0.100	-2.94	-0.101	-2.05	-0.100	-2.12	-0.003
July	-0.142	-4.16	-0.144	-2.93	-0.140	-2.96	-0.055
August	-0.213	-6.23	-0.218	-4.43	-0.209	-4.40	-0.136
September	*-0.059*	*-1.72*	*-0.059*	*-1.20*	*-0.059*	*-1.24*	-0.009
October	*-0.031*	*-1.09*	*-0.016*	*-0.40*	*-0.045*	*-1.14*	0.504
December	0.177	4.78	0.172	3.26	0.182	3.51	-0.131
Time	-3.59E-06	-6.35	-3.55E-06	-4.44	3.62E-06	-4.53	0.062
Public Holidays	-0.202	-31.77	-0.198	-21.95	-0.206	-22.97	0.664
Christmas Holidays	-0.618	-41.12	-0.298	-12.46	-0.268	-11.31	-0.897
New-year Holidays	-0.283	-16.80	-0.636	-29.46	-0.601	-28.72	-1.167
Population density	7.52E-05	97.79	7.5E-05	69.03	7.55E-05	69.26	-0.324
Veh per person*	-2.005	-249.0	-2.015	-176.4	-1.996	-175.8	-1.152
Constant	-13.67	-2270	-13.67	-1582	-13.67	-1628	0.410

* vehciles per head of population

Italic shows that these variables are not significant at 5 percent level.

2.6.3.5 Durbin-Watson test

Because the dataset contains cross-sectional time-series data, serial correlation could exist in the data: if it does, it would affect the estimates of standard errors and hence the *t* values of GLM. The Durbin-Watson test was carried out to investigate whether autocorrelation exists among the residuals. The presence of autocorrelation was tested in both the whole dataset and in each of the police force areas. Each police force is considered to be a member of a panel, with observations consisting of road accident data for each day from 1991 to 2005 with 5,479 observations. The formula given in equation 2-30 is used to calculate the Durbin-Watson statistics. The lower d_l and upper d_u values of Durbin-Watson statistics were obtained from Table 2.2 by using the number of observations and number of variables in the regression equation: the respective values for model 22 of d_l and d_u were 1.57 and 1.78 at the 0.05 level. The Durbin-Watson statistic was calculated for the whole dataset with an estimated value of 1.22. Because this value is less than the lower critical value of 1.57, the null hypothesis for the absence of autocorrelation among residuals was rejected. The same process was repeated for each of the police force area. Based on the obtained results of the test, the null hypothesis for the absence of autocorrelation among residuals was rejected for the 20 police forces however there were 6 police forces where the null hypothesis for the absence of autocorrelation was accepted. There were 25 police forces for which the null hypothesis was neither rejected nor accepted. The results are shown in Table 2.14 which is ordered by Durbin-Watson statistic so that the results are in bands.

2.6.3.6 Preferred model

Model 22 with negative binomial error structure was preferred over all other models based on the assessment of model performance as discussed in section 2.5.4. Initially the *BIC* values of all the models were compared. From section 2.6.3.1.1, model 22 was identified as having good *BIC* values. Test of the multicollinearity in section 2.6.3.3 also showed that explanatory variables used in model 22 had acceptable VIF values. It was observed that when variables of population density, vehicles per person, vehicles per road length, vehicles per surface area and ratio of road class to total road length were used together in model 29-33, it resulted in improvement of *BIC*, but these circumstantial variables had high VIFs, as a result these models were not preferred. Model 15 was also not preferred despite having better *BIC* values than model 22 because it had a Police force specific factor (51 degrees of freedom) which

subsumed the explanatory function of all area-specific variables. However, our preference was to select a model with circumstantial variables so that the inferences drawn form these parameters can be used from policy perspective. Model 27 which also had better *BIC* values was not preferred over model 22 as it resulted in difficulties in interpretation of coefficients when population density, vehicle per person and vehicle per road length were used together. From the results of the analysis of temporal effects presented in appendix Table A2.7, this showed that no substantial systematic temporal trend remains that can be represented by further quadratic temporal terms in the model and it has been represented adequately by model 22. Split sample analysis carried out on model 22 in section 2.6.3.4 showed that the estimated coefficients of model 22 are consistent and reliable.

The Durbin-Watson Test results showed the presence of serial correlation into the residuals of model 22 (Table 2.14). Due to this, Generalized Estimation Equation (GEE) with autoregressive (AR1) error term for model 22 was preferred over the GLM because it can accommodate the presence of serial correlation.

In the next section the coefficients of model 22 with GEE-AR1 and GLM with negative binomial are compared informally to identify the extent to which significance levels of the coefficients have changed.

Table 2.14: Durbin-Watson test results for Dataset 2

S.No	Police Force	DW	S.No	Police Force	DW
1	West Midlands	0.81+	27	Northumbria	1.63*
2	Essex	0.82+	28	Devon and Cornwall	1.66*
3	Metropolitan Police	0.87+	29	Durham	1.67*
4	Fife	0.92+	30	Suffolk	1.67*
5	City of London	1.08+	31	West Mercia	1.67*
6	Grampian	1.13+	32	North Yorkshire	1.68*
7	Gwent	1.24+	33	Northamptonshire	1.68*
8	Cambridgeshire	1.30+	34	Lancashire	1.68*
9	Strathclyde	1.33+	35	Warwickshire	1.69*
10	Avon and Somerset	1.34+	36	Tayside	1.69*
11	Sussex	1.36+	37	Nottinghamshire	1.69*
12	Greater Manchester	1.42+	38	Bedfordshire	1.70*
13	South Wales	1.43+	39	Humberside	1.72*
14	Central	1.44+	40	Leicestershire	1.73*
15	Cleveland	1.46+	41	Kent	1.73*
16	Cheshire	1.48+	42	Lincolnshire	1.75*
17	Merseyside	1.49+	43	North Wales	1.76*
18	West Yorkshire	1.49+	44	Dyfed-Powys	1.77*
19	Staffordshire	1.50+	45	Dumfries and Galloway	1.77*
20	South Yorkshire	1.57+	46	Dorset	1.80**
21	Thames Valley	1.58*	47	Gloucestershire	1.80**
22	Surrey	1.59*	48	Lothian and Borders	1.80**
23	Hertfordshire	1.61*	49	Wiltshire	1.81**
24	Hampshire	1.61*	50	Derbyshire	1.81**
25	Norfolk	1.62*	51	Cumbria	1.84**
26	Northern	1.62*			

+ Positive autocorrelation detected as statistically significant

**Police Forces where the null hypothesis for the absence of autocorrelation is neither accepted nor rejected*

*** Police Forces where the null hypothesis for the absence of autocorrelation is accepted*

2.6.3.6.1 Comparison of coefficients for Dataset 2

In addition to all the variables used for dataset 1, a few circumstantial variables describing the characteristics of a geographical area were also used to model dataset 2. Finally, model 22 was preferred which had population density and vehicles per head of population along with other variables of weekday 3, season, interaction of weekday 3 and season, month, time, Public holidays, Christmas and New-Year holidays. It also had an adjusted distance travelled in offset as explained in section 2.5.3 which takes account of the variations in distance travelled by day of week, month and police force. A comparison was carried out between the coefficients and *t* values obtained by GEE-AR1 and GLM with negative binomial regression as shown in Table 2.15. Because the coefficients of these two models are estimated by using the same data, they are not mutually independent so it is not possible to test rigorously for differences between them. Due to this informally the signs, magnitude and standard errors of variables were compared to identify any changes. It was found that sign for each of the coefficients was same in GEE-AR1 and GLM models except for Winter, February and September. The estimated coefficients of these variables were found to be non-significant in both models so they were not a cause of great concern. The coefficient of Summer became non-significant when GEE-AR1 was used. It was also observed that the *t* values of the weekday, Sunday, interaction variables and Public holidays increased in GEE-AR1 whereas the *t* values for the month, time, Christmas holidays, New-year holidays, population density and vehicle per person decreased. These changes are due to the presence of serial correlation in the data which is represented in the GEE model through the AR1 error structure.

From the coefficients shown in Table 2.15, the coefficient of weekday, Saturday and Sunday were 0.168, -0.028 and -0.14 respectively. This indicates greater risk of road accident per unit of travel on weekday whereas Sunday had the lowest risk per unit of distance travelled. The combined effect of month and season showed that November (0.12) had highest risk whereas August (-0.13) has the lowest risk per unit of distance travelled. Among the interaction variables, which all had significant *t* values, the coefficient of Sunday-summer (0.054) had greatest increasing effect while Sunday-winter (-0.058) had greatest reduction effect on the risk per unit of travel. These represent respectively an increase and decrease in risk of about 5 percent. The coefficient of time had negative sign and coefficient value of (-0.00000412) which indicates that risk per unit of distance travel is decreasing by 1.5 percent annually. The coefficient of Public holiday, Christmas holiday and New-year had a value of -0.185, -0.475

and -0.047 respectively which represents the variations in frequencies of road accidents occurrence on these days rather than risk. Among the other coefficients it was found that vehicles per person had a negative sign suggesting that police force areas with higher vehicle ownership per person have smaller risk of road accident per unit of travel. Population density had positive coefficient which indicates that the risk of having road accident per unit of distance travelled is greater in areas that have greater population density.

After this the combined effects of weekday 3, season, interaction of weekday 3 and season, and month were identified. Figure 2.14 shows the comparison of the risk per unit of distance travel on weekday, Saturday and Sunday by month of year, it revealed the same trend as shown in Figure 2.8 and discussed in detail in section 2.6.2.6.1. Briefly it shows that risk per unit of travel on weekdays varies substantially through the year. Greatest risk is associated with weekdays in winter and autumn. Saturdays in winter have more risk than Saturdays of other months particularly they have greater risk than some of the weekdays in spring and summer. Sunday carried the lowest risk per unit of travel than all others and this varied relatively little through the year.

Figure 2.14: Comparsion of risk per unit of distance travelled on Weekday, Saturday and Sunday by month of year (Dataset 2)

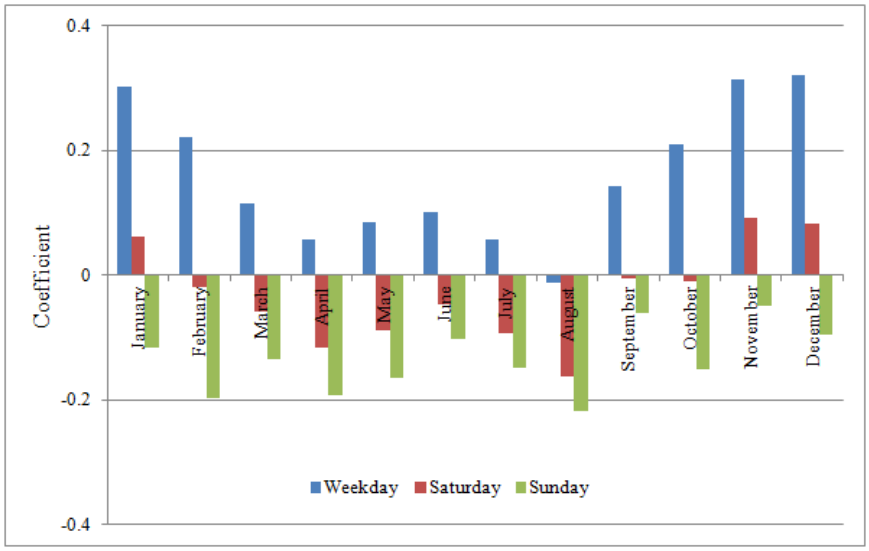


Table 2.15: Comparison of coefficient and *t* values of GEE-AR1 and GLM-Model 22-NB for coefficient validation (Dataset 2)

Variables	Comparison of models			
	Model 22-GEE-NB(AR1)		Model 22-GLM-NB	
	Coefficient	*t* value	Coefficient	*t* value
Weekday	0.168	150.16	0.164	135.54
Sunday	-0.140	-112.25	-0.141	-82.33
Summer	*0.019*	*0.61*	0.085	2.20
Autumn	0.046	2.26	0.049	2.00
Winter	*0.018*	*0.65*	*-0.057*	*-1.67*
Weekday-Summer	-0.050	-29.14	-0.049	-26.01
Sunday-Summer	0.054	28.02	0.053	20.33
Weekday-Autumn	0.026	12.28	0.024	10.48
Sunday-Autumn	-0.028	-11.41	-0.028	-8.57
Weekday-Winter	0.051	26.50	0.053	25.13
Sunday-Winter	-0.058	-26.53	-0.055	-18.45
January	0.064	2.06	0.145	3.91
February	*-0.017*	*-0.55*	*0.060*	*1.61*
March	0.058	9.78	0.050	12.88
May	0.027	4.67	0.029	7.52
June	*-0.036*	*-1.25*	-0.100	-2.94
July	-0.080	-2.81	-0.142	-4.16
August	-0.150	-5.27	-0.213	-6.23
September	*0.006*	*0.21*	*-0.059*	*-1.72*
October	*-0.030*	*-1.25*	*-0.031*	*-1.09*
December	0.084	2.74	0.177	4.78
Time	-4.12E-06	-4.43	-3.59E-06	-6.35
Public Holidays	-0.185	-34.02	-0.202	-31.77
Christmas Holidays	-0.475	-34.06	-0.618	-41.12
New-year Holidays	-0.047	-3.58	-0.283	-16.80
Population density	7.58E-05	60.50	7.52E-05	97.79
Veh per person	-2.004	-145.81	-2.005	-249.0
Constant	-13.669	-1859.40	-13.67	-2270

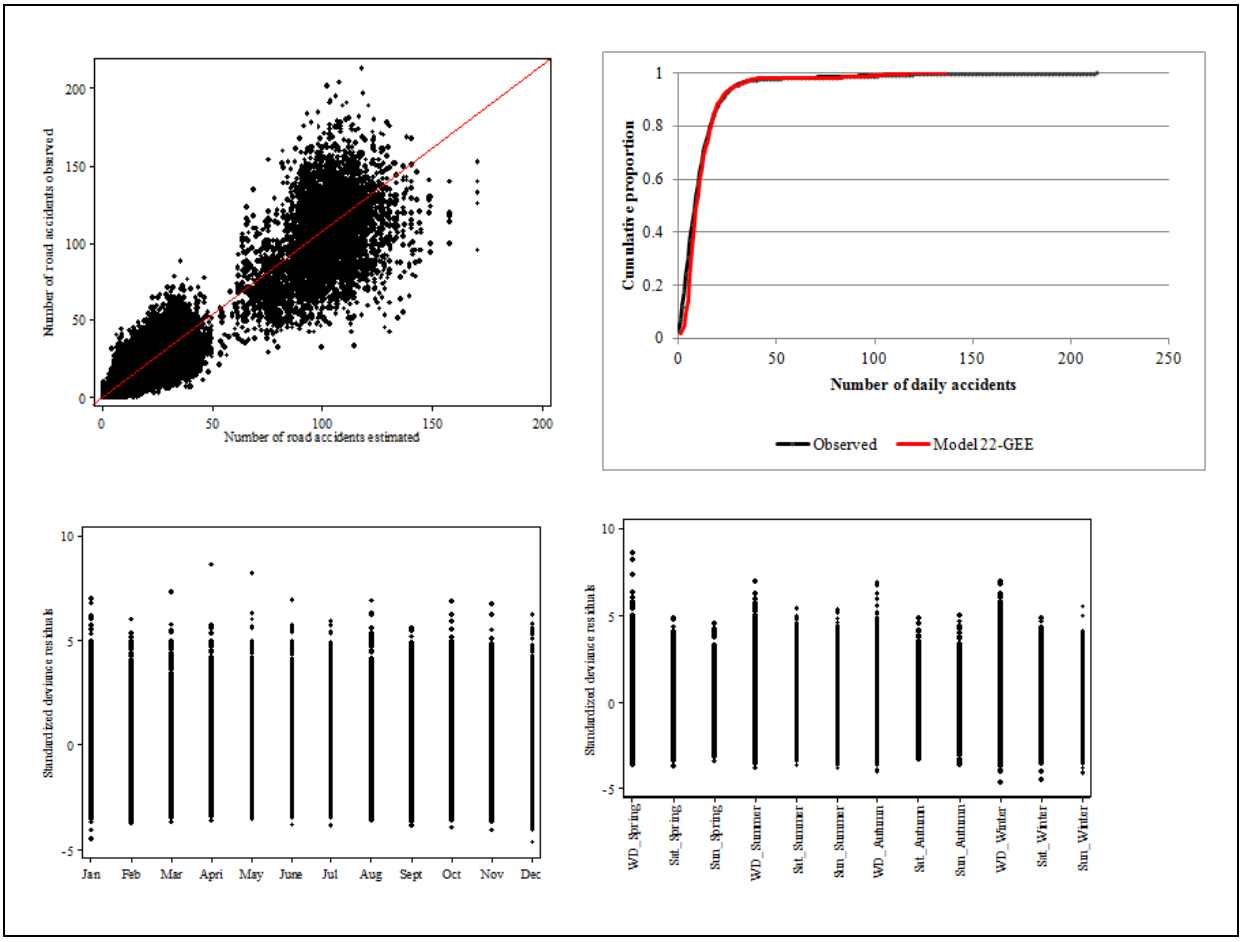
Italic shows that these variables are not significant at 5 percent level.

2.6.3.6.2 Comparison of number of road accidents observed and estimated, standardized deviance residuals and cumulative proportion graphs

Graphs for the GEE model 22 with negative binomial are shown in Figure 2.15. From the graph of road accidents observed and estimated it is observed that there are two groups present in the estimated values which are clearly visible to each side of 60. A detailed investigation was carried out to identify the characteristics of these two groups. It was observed that the Metropolitan Police Force was noticeably different from all other police forces which had a high number of road accidents on each day. The total number of observations in the dataset was 279,429 out of which 98 percent (273,729) had fewer than 50 road accidents. From the remaining 5,700 observations which had road accidents for each day greater than 50, 95 percent (5,389) belonged to the Metropolitan Police Force. This police force has only 90 observations (from 5479 observations) where number of road accidents was less than 50. These numbers clearly show that the second group of data in the graph is related to the Metropolitan Police Force which had a higher number of road accidents occurring on each day. Table A2.8 in the Appendix shows the detailed distribution of the data.

The cumulative proportion graph shows the GEE-AR1 model did not provide a precise estimate when the number of road accidents was greater than 135. However, the proportion of these observations is very small as shown in appendix Table A2.8. From the graph of standardized deviance residuals it is observed that most of the observations' standardized deviance residuals (SDR) lie in the range -5 and +5. The highest positive SDR observed was for 30 April (Friday) 1999 which is followed by 16 May (Thursday) 1991, both of the observations belonged to the City of London Police Force. Generally the SDRs for all the month lies in same range except March-June have few positive outliers. It was also found that weekdays in each of the season have higher SDRs than Saturday or Sunday when compared with the same season.

Figure 2.15: Number of accidents observed and estimated, standardized deviance residuals (Dataset 2)



2.6.3.6.3 Final model checking

Some graphs were plotted in Figure 2.16 to check visually if any problem existed in the model 22 with GEE-AR1 error structure. In the first of these graphs the deviance residuals are plotted against fitted values. The plot does not show any trend. Attention was paid to identify any increase in the deviance with increase in the predicted values, which would be cause of concern. This graph shows that the model is correct as the deviance is scattered evenly around the zero line. However, the two groups of data are clearly visible. Most of the higher number of road accident predicted values belong to the Metropolitan Police Force, which have the same level of residual deviance as other police forces. It is also observed that there were some observations with predicted values near zero. Upon further investigation it was found that most of those observations belong to the City of London and the Dumfries and Galloway police forces. There were a total of 7,213 observations where the number of road accidents observed was zero. Of these 2,173 belonged to City of London Police Force while

1,645 belonged to Dumfries and Galloway, together making up just over half of the total such observations. The details of this distribution are given in Appendix Table A2.9.

In the second graph, a normal quantile plot of standardized deviance residuals was plotted. This was used as a diagnostic tool to check that the deviance residuals have a distribution close to normal. From the graph it is shown that the quantile plot follows a straight line up to about 2.5, which supports the assumption of normality of the residuals. However beyond 2.5 the residuals deviate from the reference line which suggests that the data distribution has a longer tail at that end.

In the third graph, plotting the square root of SDR against the fitted value also did not show any noticeable pattern. The last graph of Cook's distance shows that most of the observations that had a higher peak, took place in January probably due to new-year holiday. However, Cook's distance value for those observations is in a range less than 0.002, which is substantially less than the value of 1.0 that would cause concern.

In order to verify the assumption of homoscedasticity (equal variance) in the residuals of the model 22 GEE-AR1, two different tests were carried out to identify the presence of heteroscedasticity in the residuals. The results of each of the Park and Glejser test showed that the regression of the square of the residuals on the estimated number of road accidents on each day by police force was found to be significant (Park test) and similarly for the absolute values of residuals (Glejser Test). These results shown in appendix A2.10 suggest that heteroscedasticity is present in the residuals. According to Gujarati (2009) due to the violation of the assumption of constant variance the estimated parameters are not best linear unbiased estimators (BLUE). Heteroscedasticity does not affect the unbiasedness and consistency properties of the estimators but these estimators are no longer minimum variance or efficient and the estimated standard errors are not reliable. In order to estimate the efficient standard errors for this study White's robust procedure was applied using STATA. We note that the hierarchical generalized linear model (HGLM) introduced and used in Chapter 5 allows to model variations in dispersion.

The results after applying White's procedure to model 22 GEE-AR1 with negative binomial are given in Table 2.16. This shows that the standard errors of all the variables have increased except each of March and May. However, the coefficients of Autumn, January and Time

turned to be non-significant after implementing White's corrections to standard errors. This suggests that heteroscedasticity does affect the standard errors of estimates in model 22, though it does not have a profound effect on the model structure.

Figure 2.16: Diagnostic plots for model 22 (Dataset 2)

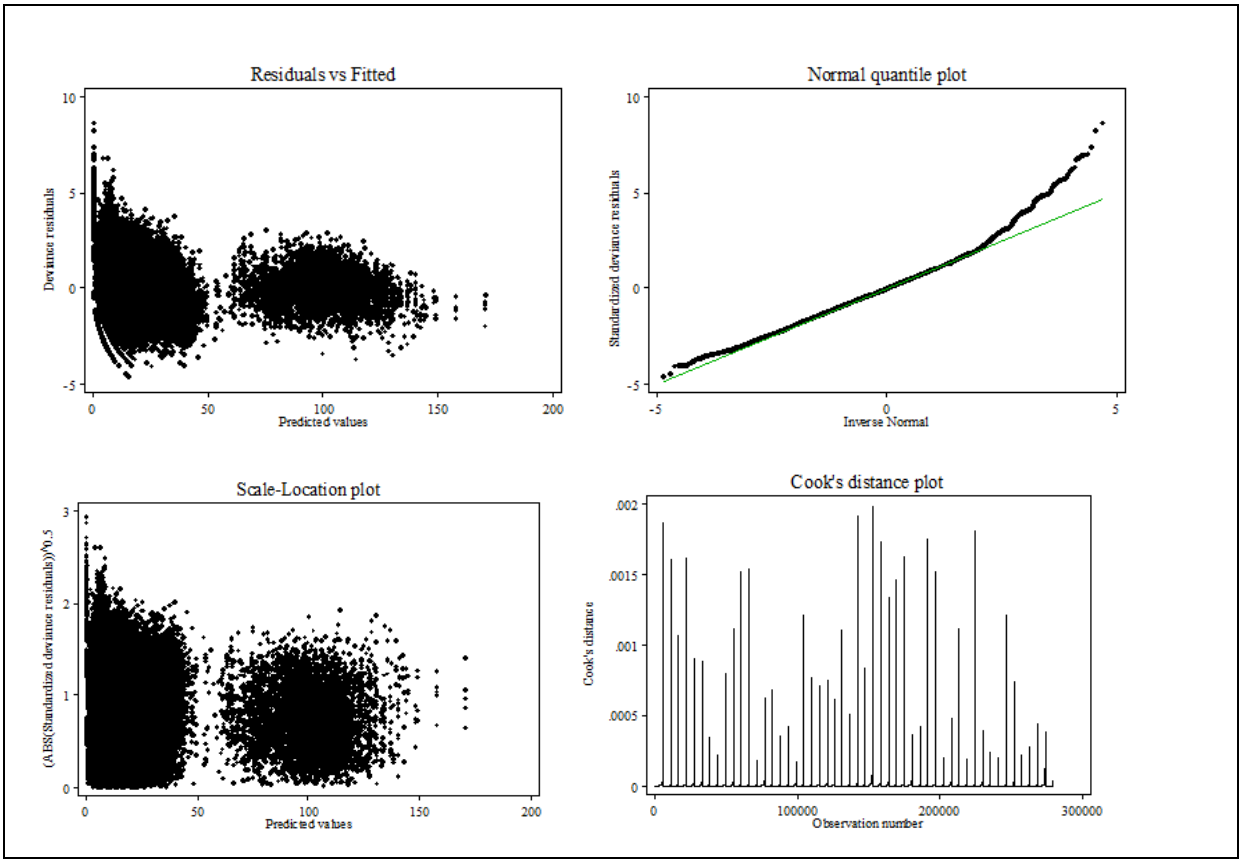


Table 2.16: Comparison of coefficient and *t* values of model 22 GEE-AR1 negative binomial after using correction for the presence of heteroscedasticity

Variables	Comparison of results of model 22-GEE-AR1			
	Before applying any corrections		White's Robust Standard Errors	
	Coefficient	*t* value	Coefficient	*t* value
Weekday	0.168	150.16	0.168	28.99
Sunday	-0.140	-112.25	-0.140	-25.18
Summer	*0.019*	*0.61*	*0.019*	*0.48*
Autumn	0.046	2.26	*0.046*	*1.81*
Winter	*0.018*	*0.65*	*0.018*	*0.56*
Weekday-Summer	-0.050	-29.14	-0.050	-12.74
Sunday-Summer	0.054	28.02	0.054	13.80
Weekday-Autumn	0.026	12.28	0.026	9.65
Sunday-Autumn	-0.028	-11.41	-0.028	-8.33
Weekday-Winter	0.051	26.50	0.051	12.98
Sunday-Winter	-0.058	-26.53	-0.058	-12.71
January	0.064	2.06	*0.064*	*1.75*
February	*-0.017*	*-0.55*	*-0.017*	*-0.45*
March	0.058	9.78	0.058	12.22
May	0.027	4.67	0.027	5.73
June	*-0.036*	*-1.25*	*-0.036*	*-1.01*
July	-0.080	-2.81	-0.080	-2.24
August	-0.150	-5.27	-0.150	-4.18
September	*0.006*	*0.21*	*0.006*	*0.17*
October	*-0.030*	*-1.25*	*-0.030*	*-1.00*
December	0.084	2.74	0.084	2.35
Time	-4.12E-06	-4.43	-4.12E-06	*-0.19*
Public Holidays	-0.185	-34.02	-0.185	-10.94
Christmas Holidays	-0.475	-34.06	-0.475	-18.50
New-year Holidays	-0.047	-3.58	-0.047	-2.08
Population density	7.58E-05	60.50	7.58E-05	3.36
Veh per person	-2.004	-145.81	-2.004	-4.29
Constant	-13.669	-1859.40	-13.669	-72.99

Italic shows that these variables are not significant at 5 percent level.

2.7 CONCLUSION:

The purpose of the analysis presented in this chapter is to formulate models for the number of road accidents occurring on each day in Great Britain. The negative binomial regression model was selected because the data were found to be over-dispersed relative to a Poisson process. A Generalized Estimation Equation (GEE) with autoregressive error terms of order 1 was preferred, because of the presence of serial correlation in the data. The offset that was adopted is the logarithm of the vehicle kilometres travelled on each day. This was based on an estimate of annual average daily traffic adjusted to take account of variations in distance travelled by each of day of week and month, so that the remainder of the model represents the risk per vehicle-kilometre of travel. A further objective was to identify the factors associated with variations in the risk of road accident occurrences. In general, the most powerful variables were found to be weekday, Saturday and Sunday. Other variables for Season, month, interaction of season and month, Public holidays, Christmas holiday, New-Year holiday, distance travelled per vehicle, population density and vehicles per person also greatly improved the performance of model.

From the estimated coefficients of the model it was found that Weekdays have greater risk per distance travelled than other days. Sunday had the lowest risk per unit of distance travelled. The interaction variable of Sunday-summer had the greatest increasing impact whereas Sunday-winter had the greatest reduction effect on the risk per unit of travel than other interaction variables. Among months of year November had the greatest risk while August had the lowest risk per unit of distance travelled. It was found that Christmas, New-Year, and other holidays have coefficients with a negative sign which shows a lower number of road accidents occurring on these days, though it was not possible to assess risk on these days because no corrections are available for distance travelled. The time variable had a negative coefficient which indicates that road accident risk is declining. It was also concluded that an increase in the distance travelled per vehicle is associated with an increase in the risk per vehicle-kilometre of being involved in road accident. Travel in Police forces areas with a higher population density have a greater risk per unit of distance travelled of road accident involvement whereas travel in police forces with greater number of vehicles per head of population will have smaller risk of road accident involvement.

Analysis of the statistical model results revealed further associations which suggest that winter and autumn are associated with more risk per unit of distance travelled in comparison to spring and summer. The risk per unit of travel on weekdays varies substantially through the year. Greatest risk is associated with weekdays in winter and autumn. Saturdays in winter have particularly more risk than Saturdays of other seasons and these Saturdays have greater risk than some of the weekdays in spring and summer. Sunday carried the lowest risk per unit of travel than all others and this varied relatively little through the year. This variation in risk per unit of travel is possibly due to change in driving behaviour and weather during these periods.

3. EFFECTS OF METEOROLOGICAL FACTORS ON ROAD ACCIDENTS

3.1 INTRODUCTION

It is widely accepted that weather plays an important role in road accidents due to rain, temperature, bad visibility, and other adverse conditions. In a recent study conducted by Norwich Union (2006), British motor claims and road accident information according to weather conditions for 2004 -2005 were examined. It was found that the amount of rainfall was a strong predictor for the number of road accidents. On a rainy day 40 percent more road accidents occur than on a complete dry day, with increased chance of multiple collisions. The research revealed extreme weather conditions of any kind could lead to an increase in the number of road accidents.

In Great Britain, a weather conditions category was first included in STATS 19 data in 1969. The information concerning weather conditions at the time of a road accident is recorded by the police officer according to nine different categories as shown in Table 3.1. From the analysis of yearly STATS 19 data (1991-2005) it was found that road accidents which occurred in fine weather without any high winds were about 77 to 87 percent of the total annual road accidents. The percentage of road accidents when raining without high winds varied from 10 to 15 percent, however all other weather conditions made a minor contribution to the total number of road accidents. The average percentage of road accidents from 1991 to 2005 by the nine weather conditions recorded in STATS 19 data is shown in Table 3.1. It should be noted that these weather conditions do not occur with equal frequency and that they might affect the traffic flows. However, they show under which weather conditions as recorded in STATS 19, more or fewer road accidents occur.

In the statistical analysis presented in Chapter 2, circumstantial variables were used in the models to characterise the area. We now hypothesis that the number of road accidents are also related to the meteorological conditions. These vary among the regions of Great Britain. It was also found in Chapter 2 that some months have more road accidents than others. Meteorological factors also vary by month. In order to investigate this variability further, this study investigated the effect of meteorological factors on road accidents whilst making allowances for different weather conditions existing across both police forces and months of the year, which were not considered in Chapter 2 as the meteorological data for all the Police

forces was not available. Without the inclusion of weather-related variables in the models, it is usually hard to explain the regional differences in safety performance. As noted above, these different weather conditions do not occur with the same frequency and may also affect traffic flows.

Table 3.1: Average percentage of road accidents occurring in different weather conditions (1991-2005)

S.No	Weather condition	%	S.No	Weather condition	%
1	Fine without high winds	79.8	6	Snowing with high winds	0.13
2	Raining without high winds	13.8	7	Fog or mist - if a hazard	0.73
3	Snowing without high winds	0.49	8	Other	1.62
4	Fine with high winds	1.2	9	Unknown	0.99
5	Raining with high winds	1.2			

Source of data: Department for Transport (2011)

This study has following objectives:

- To assess the effect of weather conditions on road accident frequency;
- To investigate the variability in number of road accidents among the months that remains even after accounting for the associated variations in weather conditions; and
- To investigate the performance of models after adding meteorological factors in addition to the circumstantial variables used in Chapter 2.

This chapter is organized as follows. Section 3.2 reviews the literature about the effects of meteorological variables on number of road accidents. Section 3.3 briefly describes the data used for this study. Section 3.4 briefly analyses the data. Section 3.5 presents the process of model development and basic structure of the model. Section 3.6 shows the model selection process, results of developed models, goodness of fit and model checks. Finally, some concluding remarks are given in section 3.7.

3.2 LITERATURE REVIEW

It has been recognized that road accidents are a consequence of the combined effects of behavioural, technological, and environmental factors. Various studies have been carried out

to determine the effects of weather on accident frequency (Brijis et al, 2008; Andrey and Olley, 1990; Codling, 1974; Edwards, 1996; Palutikof, 1991), with the understanding that weather may not be the principal cause of road accidents but it is the important environmental contributing factor. Bertness (1980) and Smith (1982) suggested that the number of road accident increases in wet weather because road users take their cars instead of walking or using public transport, thus increasing exposure but it may show a decrease in snow, with drivers either taking more care in their driving or cancelling their journeys altogether.

Weather plays a large part in the determination of road accident numbers, as a result of variation in surface condition of the road, friction, and visibility. Many theoretical and common sense reasons can be offered to explain why rain can be hazardous to traffic. The friction between the road surface and the tyres of a vehicle is reduced on a wet surface, which requires greater stopping distance. The surface on curves also becomes more slippery. Visibility at night may also be reduced due to glare and distraction of wet shining surfaces. Therefore, it is easier for a driver to lose control of a vehicle in rainy weather than in bright weather (OECD, 1976; Barzelay and Lacy, 1984).

Researchers have reported a range of increases in road accidents in rainy conditions: by 6 percent (Brotsky and Hakkert 1988), 22 percent (Smith 1982) and 52 percent (Codling 1974). Satterthwaite (1976) reported that rainy days experienced double the accident rate of dry days and Campbell (1971) showed accident rates on wet versus dry surfaces were 2.2 times higher. Brotsky and Hakkert (1988) found that on wet road days, the accident risk was 3 times greater than dry road days. Codling (1974) and Smith (1982) respectively found that 31 and 44 percent of all injury accidents occurred on rainy days. Haghighi-Talab (1973) and Bertness (1980) found that the effects of falling rain were greatest in urban areas but that road accidents were more serious in less densely settled localities where vehicle speeds are generally higher.

The medical literature provides a long well-documented list of physiological functions observed to be influenced by various meteorological phenomena. High temperature in particular is found to link to irritability and to an increase in fatigue (Boyanowski et al, 1981-82). Experiments show that in hot conditions mental performance decreases and reaction time increases (Wener and Hutchison, 1945). Among these, loss of concentration or alertness caused by heat is most likely to increase the probability of road accidents as it increases

reaction time. Thus, with increase in temperature, those making long trips along unstimulating straight roads become bored and tend to fall asleep. Temperature is the modifier of road accidents rather than the root cause (de Freitas, 1975). The seasonal pattern of increased road accidents with decreasing temperature in winter can be attributed to snow and freezing rain. The reverse in summer is in accordance with aspects of the concept of a thermal comfort range. When temperatures are beyond this range, those driving in air-conditioned vehicles may display less good judgement and hence longer reaction times. Despite this rationale, evidence linking high temperatures and road accidents is sparse. In Great Britain, Edwards (1993) used the number of road accidents that occurred each month together with the meteorological information recorded in the STATS 19 data rather than independent meteorological data to identify some relationships. She used linear regression to model the number of road accidents for each month of year. Although the conclusion drawn from this may not be reliable as the presence of over-dispersion and serial correlation were not taken into account, this does provide a starting point to use STATS 19 data for modelling road accident occurrence at the national level. Various studies were conducted to relate number of road accidents with meteorological data, some of which are summarised as follows:

Brijis et al (2008) used a Poisson Integer autoregressive model (INAR) for daily car accident data, meteorological data, and traffic flow data from the Netherlands to examine the risk effect of weather conditions on number of road accidents. Three cities Utrecht, Dordrecht, and Haarlemmermeer in the Netherlands were selected based on their proximity to national weather stations. Data for 2001 relating to traffic exposure, wind, temperature, sunshine, precipitation, air pressure, and visibility were used. From the results, they found that intensity of rain (which is the ratio between the daily precipitation amount and daily precipitation duration) and precipitation duration are highly significant variables. A positive relationship was found between the number of hours of rainfall per day and number of road crashes. Additionally, a negative, highly significant and non-linear relationship was found between the temperature and number of road accidents. It was found that lower temperatures relative to the base category (temperatures above $20^{o}C$) resulted in more road accidents, with temperatures below zero being most significant. The effects of other variables: sunshine hours, air pressure, wind, and visibility were found to be non-significant.

Schalkwyk (2006) used the amount of precipitation, number of rainy days, number of wet pavement days, and hours of wet pavement in accident frequency prediction models for both fatal and serious injury crashes. The Traffic Analysis Zone data for the year 2001 to 2002 from Michigan, Pima and Maricopa Counties in Arizona USA was used. Linear regression with logarithmic transformation of the dependent variable was carried out to estimate the number of road accidents. It was found that variables related to rain improved the goodness of fit. It was concluded that rain tends to affect and diminish safety in complex ways depending on rain frequency and intensity.

Andreescu and Frost (1998) analysed the effects of rain, mean temperature and snow on automobile road accidents in Montreal, Canada by using the three-year period 1990 to 1992 data. Average daily number of road accidents, daily values of temperature, humidity, precipitation, cloud cover, and wind speed were used. Regression equations were estimated both for entire three-year period and for each year. A strong positive relationship was found between the number of road accidents and the amount of snow in late winter and early spring. In summer months, the number of road accidents increased with rainfall. In winter, however there were large number of road accidents at low rainfall quantities and fewer road accidents on days with large rainfall. Temperature displayed a seasonal pattern of positive relationship in summer and negative relationship in winter. This study concluded that even though the population of Montreal is accustomed to driving in snowy conditions the road accident rate continues to be highest on snow days.

Edwards (1996) carried out a study to identify the relationship between road accident severity and weather. The information of both the accident severity and weather conditions was extracted from the British STATS 19 data from 1980 to 1990. In particular, this study examined actual accident severity during adverse weather. The details of accident severity and weather conditions at the time of the accident expressed in monthly aggregations were used. Severity ratios were estimated to examine the relationship between accident severity and weather condition. Initially it was found that 80 percent of road accidents occurred in fine weather with rain accounting for further 14 percent. It was found that rain-related road accidents show a consistent and significant decrease in severity when compared with road accidents in fine weather whereas the frequency of road accidents resulting in slight injury increases during rain. No statistically significant relationship between high winds and accident severity was found. Evidence for accident severity in fog was also not conclusive.

Fridstrom et al (1995) used generalized linear Poisson regression to estimate the contributions of various factors including weather to monthly road accidents numbers in provinces of Denmark, Finland, Norway and Sweden. It was found that weather conditions have a significant effect on road accident numbers although in some cases it seems counterintuitive. Rainfall increased the road accident numbers whereas snowfall had opposite effect. The results showed that in Denmark the expected monthly number of road injury accidents decrease by an estimated 1.2 percent for each additional day of snowfall during the month. The corresponding effect for fatal road accidents was even larger. Frost also had a significant effect in reducing injury accident numbers. The snow depth variable was also used for the three countries other than Denmark. This was shown to be statistically significant in reducing the number of road injury accidents in Finland and Sweden but it has a statistically non-significant value for Norway. The effect of snow depth on fatal road accidents is statistically significant in all three countries. It was also found that sudden snowfall occurring during the winter may catch drivers sufficiently unaware to cause an increased road accident risk which was witnessed by positive but non-significant coefficients of the sudden snowfall variable. The variable of daylight also has a favourable effect on the expected number of road accidents. It was also concluded that an extra one hour of light between 7 am and 11 pm will correspond to an estimated 4 percent decrease in the expected number of road injury accidents in Norway.

Andrey and Yagar (1993) used a matched sample approach to examine the data for 169 rain events and over 15,000 road accidents in the cities of Calgary and Edmonton, Canada, by using 1979 to 1983 data. The study was based on a matched sample approach, in which crash frequencies in each city were compared with matched time periods when traffic was exposed to rainfall and when precipitation did not occur. It was found that road accident risk during rainfall conditions was 70 percent higher than in other conditions. It was also suggested that accident risk returns to normal as soon as rainfall has ended, despite the lingering effects of wet road conditions.

Stern and Zehavi (1990) examined the relationship between hot weather conditions and road accidents in Israel. Seven years' road accident data from 1979 to 1985 was used along with weather details at the time of the accidents. A discomfort index was calculated and translated into physiological terms of heat stress. It was found that during medium to high heat stress which was for 43.5 percent of the total time, 56.4 percent of the total road accidents occurred.

During medium and high levels of stress more road accidents occurred than in less severe stress conditions. It was further found that road accidents associated with hot weather were mainly labelled as down to the judgement of a single person. It was concluded that road accident risk increases with the severity of hot weather, even after accounting for traffic volume.

From the literature review presented in this section it was found that meteorological variables affect the number of road accidents and their effect varies among geographical regions. Generally it was found that rainfall, temperature and snowfall have been used widely to model the frequency of road accidents. Various techniques ranging from linear regression, generalized linear regression and matched sample approaches have been used. However, the effect of meteorological factors is found to be dependent on location and type of road accidents considered. Based on the review in this section the meteorological factors shown in section 3.3.3 were adopted for this part of study.

3.3 DATA USED

The road accident and meteorological data which were considered for use in the present study are described in the following sections:

3.3.1 Road accident data

As the meteorological data was only available for some police forces with information for the monthly values of the mean maximum temperature, mean minimum temperature, rainfall, sun shine hours and number of air frost days. Due to this limitation of the meteorological data, road accident data was also transformed into number of road accidents for each month of year. The study was limited to 17 police forces because of the availability of associated meteorological data at a meteorological station based within its geographical boundaries. The selected meteorological station and police forces are shown in Figure 3.1 and 3.2. Dataset 3 consists of 3,060 observations for road accident data from 1991 to 2005. Each observation represents the number of road accidents occurring on each month of year for the whole of a police force.

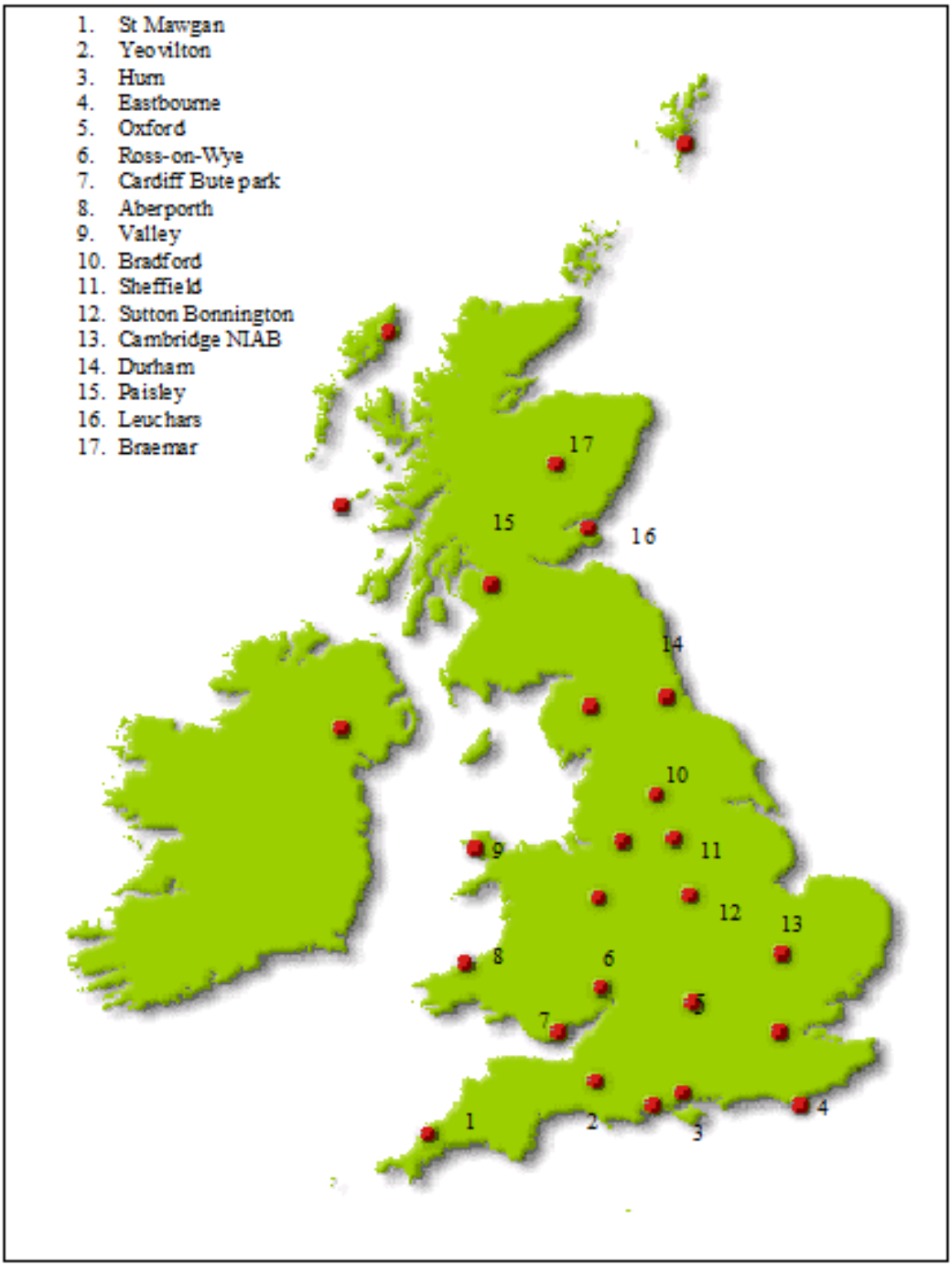
3.3.2 Meteorological data

Usually the information about snow, rainfall, and temperature is used in order to assess the effect of weather variables on road accident numbers as shown in previous studies summarised in section 3.2. The road accident data in this study is at police force level and meteorological factors may vary from place to place within each of these, their aggregation may reduce the significance of these variables in modelling road accidents at national or police force level. Due to this, all the information available from the Meteorological Office was considered with the possibility of using weather conditions jointly with the number of road accidents for police forces in Great Britain. Various meteorological datasets were made available for academic and research purposes from the Meteorological Office which are described below. Note that for the reasons explained below only historic station data was considered to be suitable for use in the present study.

3.3.2.1 Mean temperature, rainfall and sunshine data: This is a substantial dataset which gives monthly values of temperature in degrees Celsius, rain in millimetres and sunshine in hours from January 1914 to date. The data is available separately for England and Wales, Scotland, and Northern Ireland. It also gives values of temperature, rain and sunshine for each season. In this data, winter is assumed to be from December to February, spring is from March to May, summer is from June to August, and autumn is from September to November. The values of the minimum and maximum observed temperature, rainfall and sunshine for each month are also given. This data was not adopted in this study because of the aggregate nature of data as a single observation represents the whole of England and Wales.

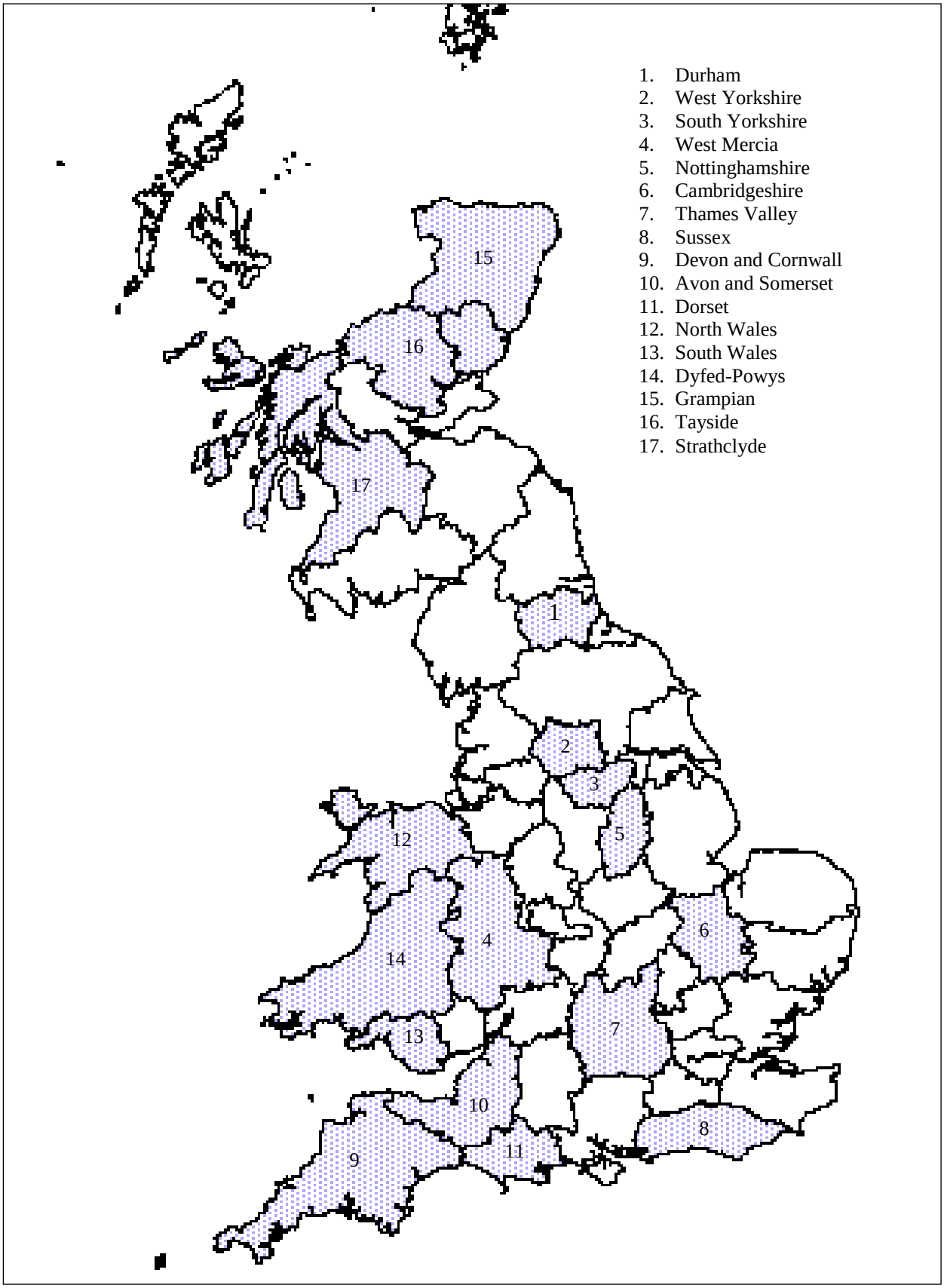
3.3.2.2 Hadley Centre Central England Temperature (HadCET): These datasets are long-period historical datasets which contain mean temperature values for each day and month of year. These daily and monthly temperatures are representative of a roughly triangular area of the United Kingdom enclosed by lines between Preston, London, and Bristol. Mean maximum and minimum temperature data are available from beginning of 1878 and are currently available free of charge. The HadCET stations are Rothamsted, Pershore, and Stonyhurst. This huge dataset was also not used for the current study because it only gave the temperature results of central England. Neither does it have any information about rainfall and sunshine.

Figure 3.1: Map showing weather stations considered for this study



Source: Meteorological office, UK (2011)

Figure 3.2: Police forces considered for this study



Source: Meteorological office, UK (2011)

3.3.2.3 UK regional precipitation series (HadUKP): The HadUKP dataset of UK regional precipitation, which incorporates the long-running England and Wales precipitation (EWP) series, begin in 1766. The precipitation values of South East England, South West England and Wales, Central England, North West England and Wales, North East England, South Scotland, North Scotland, East Scotland, and Northern Ireland were available. The information about the values of precipitation in millimetres was available in the form of daily totals, monthly totals, and seasonal totals. This dataset was also not used as it was limited only to precipitation data and it was also aggregate in nature as a single observation represented a big region.

3.3.2.4 British Atmospheric Data Centre data (BADC): This dataset which is available from BADC, UK contains land surface observations data from the Meteorological Office station network. Data of daily measurements are available for the period from 1900 to 1999. The dataset comprises daily and hourly weather measurements, hourly wind observations, maximum and minimum air temperatures, soil temperatures, sunshine duration, and hourly and daily rainfall measurements. This dataset was not adopted because data were only available up to 1999 whereas road accident data were available for the period from 1991 to 2005. In addition, information was missing for many stations and there was a lack of uniformity in the data.

3.3.2.5 Historic station data: This dataset was adopted for use in the current study. It contains observations of mean maximum and mean minimum temperature, days of air frost, total rainfall, and sunshine hours for each month of year for 25 stations across UK. Three stations were closed during the period studied: Greenwich (in 2004), Ringway (in 2004), and Southampton (in 2000) so the incomplete data from these stations was not used. The stations at Lerwick, Stornoway airport, Tiree, and Armagh were also not considered. The station at Newton Rigg which was based in the Cumbria police force area was also excluded as it did not record sunshine hours. Thus a total of 17 stations for which the data was available were selected each representing meteorological conditions in one police force. These are shown on the map in Figure 3.1. The meteorological data for these stations was extracted and used jointly with the road accident data.

3.3.3 Variables available from historic station data

The following variables from the historic station data were adopted for use in this study.

3.3.3.1 Monthly rainfall: This is the total sum of rainfall for all days of a month. Usually measurement of rainfall is made at 0900 GMT which gives the amount of rain that has fallen in previous 24 hours. The unit of monthly rainfall is the millimetre (mm). A measurement of 1 mm of rainfall indicates that if none of the rain that fell in the surrounding area had drained or evaporated away, it would have covered the entire surface to depth of 1mm.

3.3.3.2 Mean maximum monthly temperature: This is the mean of the maximum daily temperature for all the days of the month. The reading is usually made at 0900 GMT from a thermometer that has a bimetallic strip which gives a reading for the previous day. The maximum temperature usually occurs at around 1400 GMT. Temperature is measured in degrees centigrade.

3.3.3.3 Mean minimum monthly temperature: This is the mean of the minimum daily temperature for all the days of the month. The reading is usually made at 0900 GMT, always at the same time, from a thermometer that records the values of minimum and maximum temperatures. The minimum temperature usually occurs at about dawn.

3.3.3.4. Total sunshine duration: This is the sum of daily bright sunshine hours of the month. A Campbell-Stokes sunshine recorder or a Kipp and Zonen sensor are normally used to measure the daily amount of sunshine. The sunshine duration is measured in hours.

3.3.3.5. Air frost days: This is the number of days in a month when the air temperature falls below freezing. A Stephenson screen is used to measure the temperature. When the temperature within this screen reaches 0^0C there is said to be an air frost. The unit of measurement is number of days in a month on which air frost occurred.

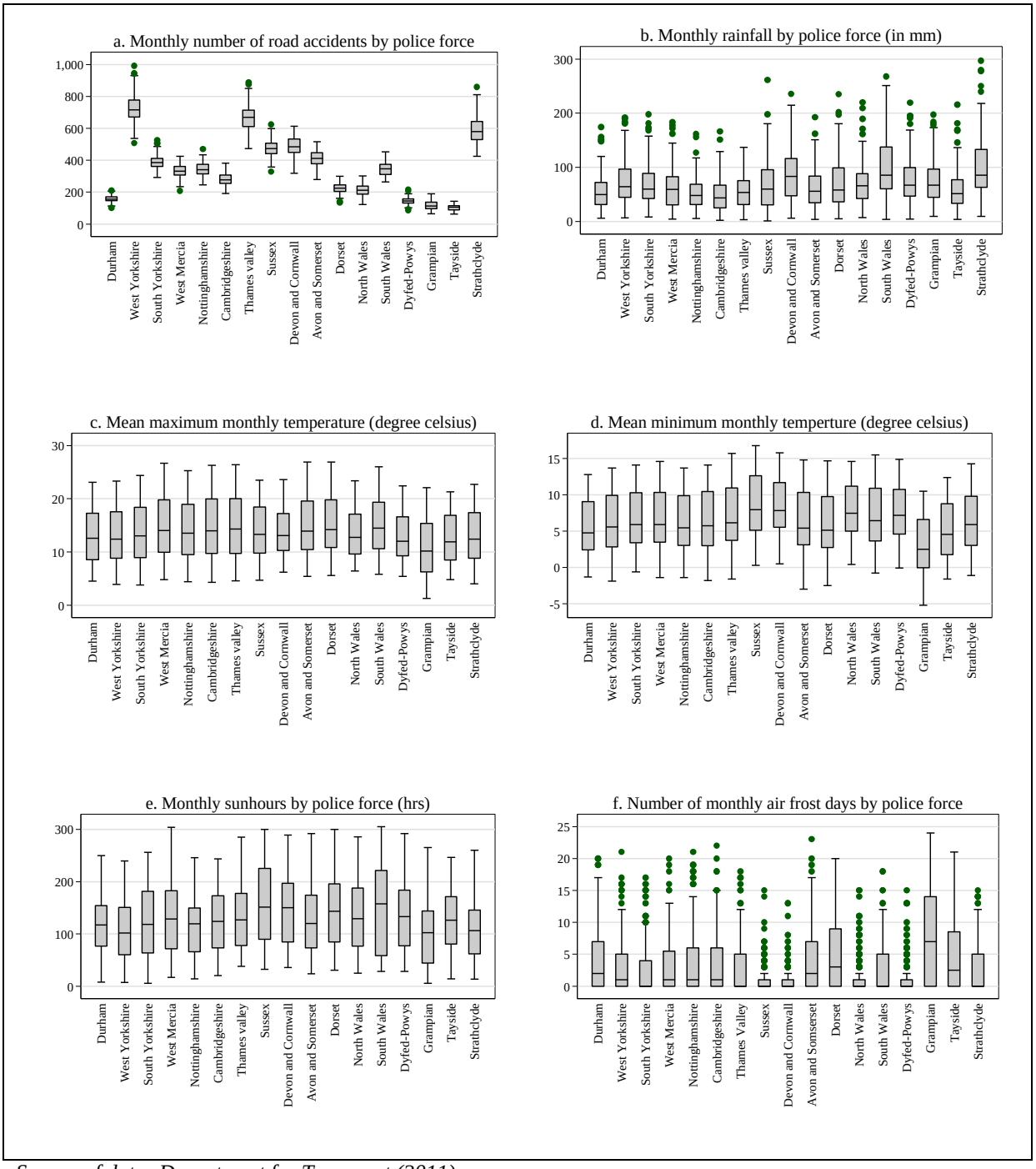
3.4 DATA ANALYSIS

The combined STATS 19 and meteorological data from 1991 to 2005 are examined by using graphical plots in Stata software.

Box plot (a) in Figure 3.3 shows that there was a substantial range of typical road accident numbers among the police forces as would be expected from their differing sizes and populations. This also shows variability from month to month within each police force area. The interquartile range of Tayside and Grampian was small in comparison to other police forces. Of the three Welsh police forces, South Wales police force had the highest number of road accidents whilst in Scotland, Strathclyde had the highest number of road accidents occurring on each month of year. From the available data it was found that the South East and South West regions of England had a greater number of road accidents than the East of England, West Midlands, and East Midland regions. Box plot (b) shows Wales had a higher amount of rain than England and Scotland. The police forces of South Wales and Strathclyde have highest amount of rainfall followed by Devon & Cornwall. The lowest rainfall was found in Cambridgeshire. Box plot (c) shows the median of the mean maximum temperature of the English police force areas was about the same for all police forces and ranged between 13 and 14 ^{o}C. In general it can be seen that the English police force regions are warmer than those in Scotland.

Box plot (d) indicates that effect of winter is not as severe, especially in Sussex and in Devon & Cornwall, as in other police forces. Scottish police force areas were found to be less warm than others. The Welsh police force areas were found to have higher mean minimum temperature than most of the English and Scottish police forces. Box plot (e) reveals that the South Wales police force had the highest number of monthly sun hours followed by Sussex and Devon & Cornwall while the lowest number of sun hours occurred in West Yorkshire. It can also be seen from the graph that the interquartile range of the South Wales force area was greater than all police forces, indicating a higher difference in sun hours between summer and winter months. Durham police force was found to be least variable. Box plot (f) shows that Scotland had a greater number of air frost days than England and Wales. It was also found that the interquartile range for the Sussex, Devon & Cornwall, North Wales, and Dyfed-Powys police force areas was smaller than others indicating that these police forces had few air frosts days with less variation in different months of the year.

Figure 3.3: Box plot of STATS 19 data (Dataset 3: 1991-2005)



Source of data: Department for Transport (2011)

From Figure 3.3 it is also found that some police forces had warmer afternoons but colder mornings as the difference between mean maximum monthly temperature and mean minimum monthly temperature varied from 4.8^0C for Dyfed-Powys to 9.05^0C for Dorset. It was also observed that some police forces that had a higher mean minimum temperature did not have a higher mean maximum temperature. For example the Sussex police force had the highest median for mean minimum temperature but it had a lower mean maximum

temperature than many other police forces. Grampian had both these temperatures at lowest level. Similarly it was observed that maximum temperature and sunshine hours carry different information. July and August were the months in all police forces when the mean of the maximum monthly temperature was high. On other hand the mean for sun hours was high for the month of May in most of the police forces, however in some police forces it was either in June or July.

It is observed that each meteorological factor represents a particular characteristic of a police force like some police forces have large difference in temperature between winter and summer months. Due to this, all the available meteorological factors were used in models to identify their impact on the road accidents occurrence on each month of year.

3.5 MODEL DEVELOPMENT

A total of 24 models were developed by using the Generalized Linear Model (GLM) with negative binomial regression using the Stata software. In the first step, a model was developed with a constant term and an appropriate offset. After this, a stepwise incremental approach was followed by adding different variables in the model. The lattice of model development is shown in Figure 3.4.

3.5.1. Variables used

The following variables were incorporated into the model to estimate the number of road accidents by each month of year within each police force.

1. Police force (17 levels)
2. Month (12 levels)
3. Season (4 levels: Spring, Summer, Autumn and Winter)
4. Time as a variate (measured in months, with values from 1 to 180, January 1991 to December 2005).
5. Population density
6. Vehicles per head of population
7. Meteorological variables

a. Mean maximum monthly temperature
b. Mean minimum monthly temperature
c. Monthly rainfall
d. Monthly sunshine hours
e. Monthly number of air frost days

Population density and Vehicles per head of population were used, as from chapter 2 it was found that they have significant effect on the number of road accidents. In the models, both maximum monthly temperature and minimum monthly temperature were also used together to account for the large variation in meteorological conditions between winter and summer months. Although the pattern of maximum and minimum temperature remain similar over the month, it was observed from the data that omitting one variable or the other may result in not considering the particular nature of certain police forces. Due to this, all meteorological variables were used in various combinations in the modelling process to represent the variation among police forces. Assessment of the model performance was based on the criteria that was discussed in section 2.5.4.

3.5.2. Basic structure of the model

In this chapter all models which were developed for Dataset 3 are shown in Figure 3.4. Each observation of the dependent variable *y* represented the number of road accidents occurring during each month of the observation period (1991-2005). Data for each month was used rather than each day due to the limited availability of the associated meteorological data for the police forces. Because the unit of observation in this dataset is month, the adjusted total distance travelled (vehicle kilometres) in each police force during each month was used as the offset. This was estimated by adjusting the annual average total distance travelled as follows.

First the annual average distance travelled was adjusted according to the calendar month (by applying month correction factors obtained from Department for Transport) to give an average distance travelled for a day of that month. This was then multiplied by the number of days in the month to give a total distance travelled during the month. Finally, this was factored according to the number of vehicles registered in each police force area during that

year. The result of this is an estimate of the distance travelled in each police force area during each month of each year. This is proportional to each of:

- Distance travelled on a typical day of that month
- Number of days in the month
- Number of vehicles registered in the police force

The remainder of the model can then be interpreted in terms of the risk of accidents per vehicle-kilometre of distance travelled in a police force area during a month.

To achieve this, the following model structure was used for Dataset 3.

$$u_{ij} = \exp\left(O_{ij} + \mathbf{x}'_{ij}\ \boldsymbol{\beta}\right) \qquad 3\text{-}1$$

where i represents the observation (time in months 1 to 180), j represents the police force (1 to 17)

u_{ij} is the estimated mean number of road accidents for each month of year.

O_{ij} is the offset calculated as $\ln\left(d_{ij}\right)$

Then $u_{ij} = d_{ij}\ \exp\left(\mathbf{x}'_{ij}\ \boldsymbol{\beta}\right)$ 3-2

where d_{ij} is the adjusted total distance travelled (vehicle kilometres) in month *i* within the police force area *j*.

The linear predictor in this model then represents the mean risk of accident involvement per unit of travel in police force area *j* during month *i*.

3.6 MODEL SELECTION PROCESS, GOODNESS OF FIT AND MODEL CHECKS

The model selection procedure described in section 2.5.4 was applied to distinguish among many available models. The results of all the developed models shown in Figure 3.4 were compared. The details of all these models and the various checks that were used to identify the appropriate model are given in section 3.6.1.1 to 3.6.1.5.

3.6.1 Model Selection Procedure

The procedure shown in section 2.5.4 was used to identify the most appropriate model out of the many developed models which can estimate the number of road accidents on each month of the year and can give some insights on the variables used in the modelling. All of the models presented here were developed using negative binomial regression. Additional meteorological variables were also included in different combinations to investigate their effect on road accident occurrence. The following section shows the results of the tests carried out for model selection:

1. In section 3.6.1.1 *BIC* values of all the models are compared to check their performance
2. In section 3.6.1.2 temporal effects were analysed to investigate any temporal effect remaining that is not captured by the models.
3. In section 3.6.1.3 variance inflation factors were used to check for the presence of multicollinearity in the data.
4. In section 3.6.1.4 split sample tests were carried out to validate the performance of the model by comparing the coefficients, deviance and log-likelihood values.
5. In section 3.6.1.5 the presence of serial correlation in the residuals was tested by using Durbin-Watson test.

3.6.1.1 Negative binomial regression model (Dataset 3)

A total of 24 models were developed with different combinations of variables as shown in Figure 3.4. The logarithm of the adjusted total distance travelled in each police force area during each month was used as the offset. In the process of model development the correction to adjust the offset to account for variations in distance travelled for each month of the year were applied from model 3 onwards only when the month variable was used as an explanatory variable. In model 4, when the simplified categorical variable of Season was used, these month adjustments to the offset were retained. The *BIC* values were calculated and used to assess the performance of these models.

Model l used only the constant term and an offset, giving *BIC* values of 35,400. Variables of police force, month and season were added individually in models 2, 3, and 4 respectively. Results showed that introducing the police force variable in model 2 improved the *BIC* by

1,186 (3 percent) by using 16 more degrees of freedom in comparison to model 1. In model 3 and 4, month of year and season were used respectively which showed that month variable performed better than season as model 3 had better *BIC*. Similarly comparison of results of models presented in appendix Table A3.1 showed that month constantly performed better than season and this preference does not diminish with increasing model complexity. In view of this, month was preferred over season and carried forward to model 5.

In model 5, time variable improved the *BIC* by 911 (2.5 percent) in comparison to model 3. Next, population density and vehicles per person variables were added in models 6 and 7. These models clearly performed better than model 2. The *BIC* value of model 6 was 34,092 showing an improvement of 1,308 (4 percent) in the *BIC* value in comparison to model 1. The vehicle per head of population performed better than population density as its addition in model 7 improved the *BIC* value by 1,195 in comparison to model 6. In model 8, the police force variable was used in addition to month, time, population density and vehicles per person. The results of model 8 were compared with model 24 to get an understanding of the improvement in the model due to the addition of meteorological variables.

The meteorological variables were introduced individually into the models from model 9 to model 13. It was found that out of all the introduced meteorological variables mean minimum monthly temperature (model 10) improved the model *BIC* value by 49 in comparison to model 7 whereas the other meteorological variables could not improve the *BIC* when used individually. After this, from model 14 onwards the meteorological variables were used in different combinations: this showed that model 14 with maximum and minimum temperature and model 15 with minimum temperature and rainfall performed better than models 16 to 23 (except model 18, 22 and 23) in terms of *BIC* value. In models 18, 22 and 23 maximum and minimum temperature, amount of rainfall, hours of sunshine and number of air frost days were used in different combinations.

The police force variable was then introduced into the model 24, after adding all meteorological variables of maximum and minimum temperature, amount of rainfall, hours of sunshine and number of air frost days to investigate whether the police force variable can subsume the remaining differences among the various geographical areas. The *BIC* value for model 24 was 31,473 which was found to be better than all other models. Comparison of model 2 with model 24 showed an improvement of 2,741 in *BIC* value for model 24 which is

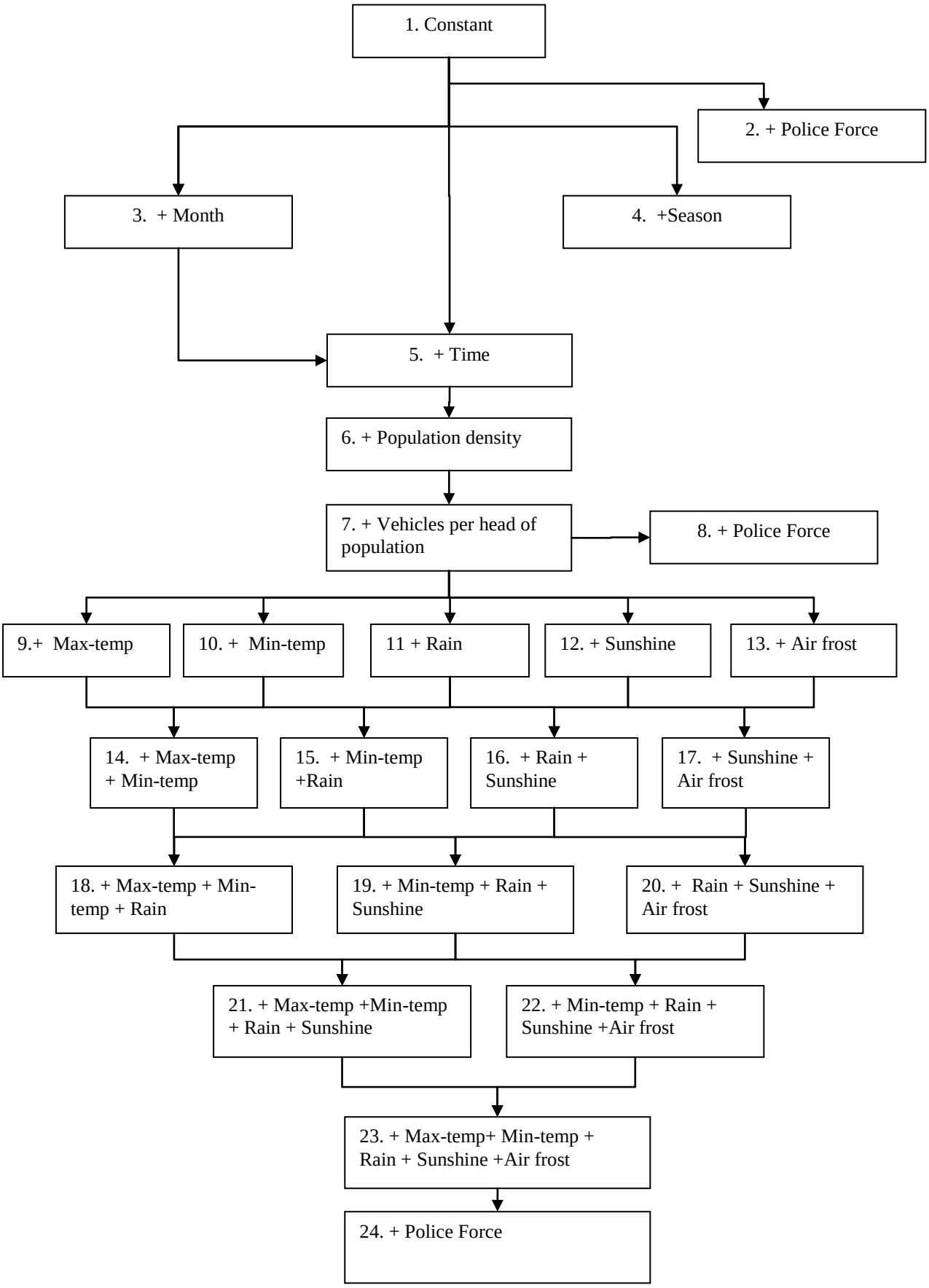
contributed by the inclusion of the variables of month, time, population density, vehicles per head of population and meteorological variables. However, it is also observed by comparing model 8 with 24, that all meteorological variables contributed an improvement of 131 in the *BIC* value: this shows that meteorological conditions within police force areas contribute to model performance. Based on the *BIC* results model 10, 14, 15, 18, 22 and 23 were considered for further analysis. Detailed results of the performance of these models are shown in Table 3.2.

Table 3.2: Results of models for the police forces with meteorological variables (Dataset 3)

Model	**D.F**	**Scale**	**Log-Likelihood**	***BIC***
1	1	0.0630	-17,696	35,400
2	17	0.0396	-17,039	34,214
3	12	0.0642	-17,725	35,547
4	4	0.0660	-17,766	35,564
5	13	0.0470	-17,266	34,636
6	14	0.0387	-16,990	34,092
7	15	0.0242	-16,388	32,897
8	31	0.0139	-15,677	31,604
9	16	0.0242	-16,387	32,902
10	16	0.0237	-16,360	32,848
11	16	0.0241	-16,385	32,899
12	16	0.0242	-16,388	32,904
13	16	0.0242	-16,387	32,903
14	17	0.0236	-16,354	32,844
15	17	0.0237	-16,359	32,855
16	17	0.0241	-16,384	32,905
17	17	0.0242	-16,387	32,910
18	18	0.0236	-16,354	32,852
19	18	0.0237	-16,359	32,863
20	18	0.0241	-16,384	32,912
21	19	0.0236	-16,353	32,858
22	19	0.0235	-16,343	32,838
23	20	0.0233	-16,338	32,836
24	36	0.0129	-15,592	31,473

BIC represents the Bayesian information criterion

Figure 3.4: Lattice of model development for Dataset 3



3.6.1.2 Analysing the temporal effects

The procedure presented in section 2.6.2.2 was used to investigate for the presence of further temporal effect that was not represented in the models. For this, time and square of time variables were added to model 1-4 whereas from model 5 onwards only square of time was added as these models already had time variable in the linear predictor. The resulting improvement in the *BIC*, coefficients and *t* values of time and square of time, and their variance inflation factors were then examined.

From the results shown in Appendix Table A3.2, an improvement of at least 800 was observed when time and square of time variables were added to each of the model 1-4. This shows that these models did not account for temporal effects. From model 5 onwards only square of time was added as an explanatory variable. The results of this show that from model 5 onwards there was no improvement in *BIC*, showing that temporal trend has been adequately represented by these models. From model 5 to 24, the square of time variable has non-significant *t* values (except model 24). Model 24 with 36 degrees of freedom which had the better *BIC* value than all other models showed that after adding square of time variable (one degree of freedom) the value of *BIC* becomes slightly less preferable which indicates that quadratic temporal trend is not required in the model.

3.6.1.3 Checking for the presence of multicollinearity:

As discussed in Chapter 2, section 2.6.2.3, the presence of multicollinearity will cause the standard errors to be inflated, the sign and magnitude of the coefficients of variables may also vary. Due to this, variation inflation factors (VIF) were estimated in order to measure the severity of collinearity and to quantify the increase in the variance of the estimated coefficients.

Table 3.3 shows the VIFs of models 9-24 where the variables of time, population density, vehicles per person, maximum monthly temperature, minimum monthly temperature, amount of rain fall, sunshine hours and number of air frost days in each month were used in different combinations. The results show that in models 9-13 in which each meteorological variable was used individually with other explanatory variables had acceptable values of VIF whereas model 9 had slightly high VIF of 9.05 (for maximum temperature) but it is still under the

critical value of 10. Results of the models 14-24 show the models that included both maximum temperature and minimum temperature (models 14, 18, 21 and 23) produced high VIF for these variables, so we conclude that these two variables are correlated. Provided that we apply the model to places where this correlation remains, multicollinearity will not cause great difficulty. However, the partial effects of maximum temperature and minimum temperature cannot be identified reliably. In model 22 minimum temperature, amount of rainfall, sun shine hours and air frost days were used but minimum temperature had high VIF of 13.21. Finally, model 24 in which all the explanatory variables are used together had high VIF, strongly suggesting the presence of multicollinearity. It is observed from Table 3.3 that in model 10 minimum temperature had a VIF of 6.15 whereas in model 15 minimum temperature and rainfall had a VIF of only 6.36 and 1.2 respectively. Table 3.3 shows the variance inflation factors of the models (9-24) for Dataset 3.

Table 3.3: Variance inflation factors of all the models (Dataset 3)

Model	Time	P.D	V/P	Max t	Min t	Rain	Sun hrs	A.F
9	1.41	1.08	1.51	9.05	-	-	-	-
10	1.41	1.06	1.53	-	6.15		-	-
11	1.41	1.04	1.46	-	-	1.16	-	-
12	1.41	1.04	1.52	-	-	-	4.51	-
13	1.41	1.05	1.45	-	-	-	-	2.03
14	1.41	1.54	1.54	14.17	9.63	-	-	-
15	1.41	1.06	1.57	-	6.36	1.2	-	-
16	1.41	1.04	1.53	-	-	-	4.15	2.03
17	1.41	1.05	1.53	-	-	1.21	4.32	-
18	1.42	1.08	1.57	14.8	10.38	1.25	-	-
19	1.41	1.06	1.61	-	6.59	1.27	4.48	-
20	1.41	1.05	1.54	-	-	1.25	4.32	2.1
21	1.42	1.08	1.61	16.34	10.41	1.29	4.94	-
22	1.41	1.06	1.62	-	13.21	1.27	4.58	4.21
23	1.42	1.08	1.62	16.38	16.53	1.29	5.07	4.22
24	2.07	2885.1	3.09	35.51	33.96	1.49	7.23	4.39

P.D=Population density, V/P=vehicles per person, Max t=mean maximum monthly temperature, Min t=mean minimum monthly temperature, Sun hrs= Monthly sunshine hours, A.F= monthly number of air frost days

3.6.1.4 Split sample tests

After analysing the *BIC*, temporal effects and *VIF* values according to the criteria discussed in section 2.5.4, model 15 was taken forward for further investigation. In order to check the consistency of model and its parameters, split sample validation tests were carried out. To this end, the whole dataset was partitioned randomly into two parts as explained in section 2.6.2.4. Each part contained 1,530 observations. The following datasets were used to cross-check and validate the results of model 15.

Full dataset = Dataset A
Dataset first portion = Dataset B
Dataset second portion = Dataset C

Stata software was used to estimate the model parameters of Dataset B and C which were then compared. The results in Table 3.4 show that the estimated value of log-likelihood for dataset B and C differed by value of 72. The optimised likelihood for dataset B and C was -8,140 and -8,212 respectively. However, the deviance of dataset B was higher only by value of 3. After this, the coefficients of dataset B and C were interchanged to estimate the number of road accidents for each month and values of log-likelihood and deviance were estimated.

After interchanging the coefficients, the log-likelihood of dataset B was estimated to be -8,152 which differed by only 12 below the optimised value for dataset B. Because the model parameters are not optimised in this case, there are 17 more degrees of freedom in the residuals: this gives a likelihood ratio test statistic of 24 on 17 degrees of freedom, which is less than the critical value of 27.59 at 0.05 significance level. Therefore the null hypothesis cannot be rejected that parameters fitted to dataset C are appropriate for dataset B. In the same way, after interchanging the coefficients, the log-likelihood of model C was estimated to be -8,222 which had a difference of only 10 below the optimised value of model C: this gives a likelihood ratio test statistic of 20 on 17 degrees of freedom, which is less than the critical value of 27.59 at 0.05 significance level. Therefore the null hypothesis cannot be rejected that parameters fitted to dataset B are appropriate for dataset C.

It is also found that the log-likelihood and total deviance values of data A were better than the sum of the two corresponding values. The log-likelihood had a difference of about 5 while

deviance was found to differ by 23. Table 3.4 shows that the log-likelihood values of all the four models were consistent and did not differ with statistical significance ($\alpha = 0.05$). These results showed that model 15 is stable and the parameter estimates are reliable.

Table 3.4: Split sample validation results for Dataset 3

Split sample validation				
		Model coefficients (k=17)		
Data		A	B	C
A		$\mathbf{x}'_A \boldsymbol{\beta}_A$		
	n	3,060		
	Likelihood	-16,359		
	Deviance	3,139		
B			$\mathbf{x}'_B \boldsymbol{\beta}_B$	$\mathbf{x}'_B \boldsymbol{\beta}_C$
	n		1,530	1,530
	Likelihood		-8,140	-8,152
	Deviance		1,571	1,594
C			$\mathbf{x}'_C \boldsymbol{\beta}_B$	$\mathbf{x}'_C \boldsymbol{\beta}_C$
	n		1,530	1,530
	Likelihood		-8,222	-8,212
	Deviance		1,589	1,568
Total	Likelihood	-16,359	-16,362	-16,364
	Deviance	3,139	3,160	3,162

k represents the number of explanatory variables in the model and n represents number of observations

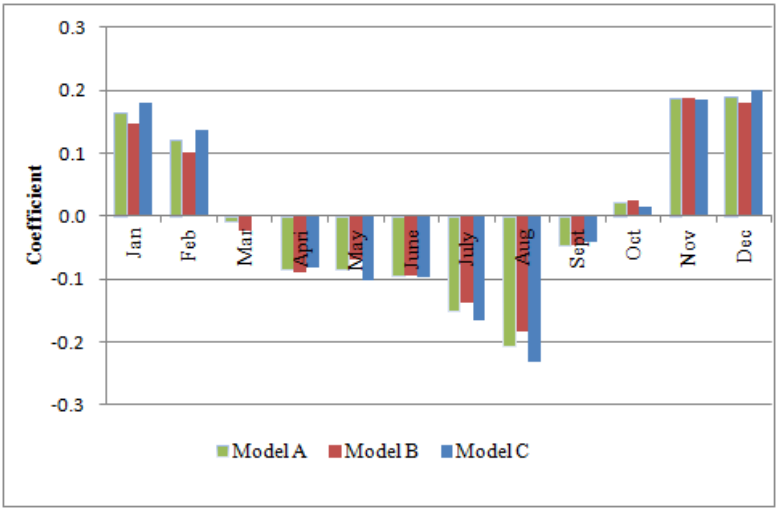
In the second step of the validation process the coefficients fitted to dataset A, B and C are compared. It is observed that coefficients of all variables and *t* values of the explanatory variables are consistent and carried the same sign in all three models except March which had non-significant *t* value in all three models. Some variables changed from significant to being non-significant variables across different models. October which had significant *t* value in model A turned to be non-significant in model B and C. The *T* test was used to compare the coefficients of dataset B and C. Formula 2-32 was used to estimate the *T* test values. It is found from the *T* test values that the coefficients of model B are not significantly different from the coefficients of model C. For all variables the estimated values of T_{BC} are less than 1.96 which suggests that coefficients have not changed significantly. The comparison of coefficients and *t* values are shown in Table 3.5 and Figure 3.5.

Table 3.5: Comparison of coefficient and *t* values of GLM-Model 15-NB for coefficient validation (Dataset 3)

Variables	Comparison of the coefficients and *t* values of the models						
	Model A		Model B		Model C		T test
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	T_{BC}
January	0.165	12.44	0.148	7.83	0.181	9.66	-1.263
February	0.120	9.14	0.103	5.64	0.137	7.22	-1.307
March	*-0.010*	*-0.85*	*-0.022*	*-1.36*	*0.001*	*0.06*	-1.012
April	-0.086	-8.20	-0.090	-6.14	-0.083	-5.52	-0.339
May	-0.084	-8.26	-0.068	-4.97	-0.102	-6.72	1.658
June	-0.096	-7.85	-0.095	-5.60	-0.097	-5.49	0.105
July	-0.152	-10.23	-0.139	-6.63	-0.164	-7.79	0.865
August	-0.208	-14.07	-0.183	-8.85	-0.231	-10.97	1.646
September	-0.047	-3.90	-0.047	-2.93	-0.042	-2.26	-0.213
October	0.021	2.10	*0.026*	*1.83*	*0.016*	*1.13*	0.502
November	0.186	17.13	0.187	12.98	0.184	11.28	0.135
December	0.190	14.53	0.180	9.83	0.200	10.68	-0.782
Time	-0.001	-15.69	-0.001	-11.29	-0.001	-10.99	0.165
Pop-density	0.0002	18.74	0.0002	14.28	0.0002	12.33	1.125
Veh/Person	-1.310	-41.00	-1.289	-27.99	-1.328	-29.83	0.614
Min-temp	0.014	7.27	0.011	4.12	0.017	6.06	-1.476
Rain	*7.4E-05*	*0.98*	*1.1F-04*	*0.99*	*3.43E-05*	*0.31*	0.440
Constant	-13.993	-817.19	-13.995	-579.75	-13.990	-573.79	-0.144

Italic shows that these variables are not significant at 5 percent level.

Figure 3.5: Comparison of coefficients of models using GLM-Negative binomial (Coefficient validation-Dataset 3)



3.6.1.5 Durbin-Watson test

The Durbin-Watson test was used to investigate the presence of serial correlation among the residuals. Each police force is considered to be a member of a panel which consists of 180 observations, each representing the number of road accident during a month from 1991 to 2005. The formula given in equation 2-30 is used to estimate the value of the Durbin-Watson statistic. The lower d_l and upper d_u critical values of the statistic were obtained from Table 2.2 by using the number of observations and number of variables in the regression equation. The respective values of d_l and d_u were 1.57 and 1.78. This table also showed the regions of the acceptance and rejection of the null hypothesis for the absence of serial correlation. The Durbin-Watson statistic was calculated for whole dataset and for each police force. The estimated value of Durbin-Watson statistic for whole dataset was 0.98 which was in the first region (less than 1.57) as a result of this, the null hypothesis for the absence of serial correlation among residuals was rejected. The same process was repeated for each police force. It was found from the results shown in Table 3.6 that null hypothesis for the absence of autocorrelation among residuals was rejected for the 16 police forces. However, the hypothesis for the absence of autocorrelation for Sussex police force was neither rejected nor accepted as the estimated value of the Durbin-Watson statistic lies between 1.57 and 1.78. Based on the estimated results shown in Table 3.6, it is concluded that serial correlation exists in the residuals as a result of which *t* values obtained by the GLM may be inflated.

Table 3.6: Durbin-Watson test results for Dataset 3

S.No	Police Force	DW	S.No	Police Force	DW
1	Durham	1.42	10	Avon and Somerset	0.53
2	West Yorkshire	1.47	11	Dorset	1.14
3	South Yorkshire	0.89	12	North Wales	0.98
4	West Mercia	1.14	13	South Wales	0.69
5	Nottinghamshire	0.87	14	Dyfed-Powys	1.17
6	Cambridgeshire	0.29	15	Grampian	0.31
7	Thames Valley	1.50	16	Tayside	1.49
8	Sussex	1.63*	17	Strathclyde	1.07
9	Devon and Cornwall	1.32			

**panel where the null hypothesis for the absence of autocorrelation is neither accepted nor rejected.*

3.6.1.6 Preferred model:

Model 15 was preferred based on the criteria presented in section 2.5.4. In section 3.6.1.1, model 15 had good *BIC* values and tests for multicollinearity showed that explanatory variables in model 15 had acceptable VIF values. Other models in which meteorological variables were used in different combinations had high VIFs for these variables, so they were not preferred despite having better *BIC* values as the true effects of these variables can not be identified correctly.

Model 15 was preferred over all other models despite some of them (Model 8, 10, 14, 18 and 22-24) having better *BIC* values (Table 3.2). It was found that the models in which mean minimum monthly tempertaure and mean maximum monthly temperature (model 14, 18 and 22-24) were used together produced high VIFs suggesting that they are collinear so these models were not preferred. Model 8 was not preferred as it had a specific factor for Police force (17 degrees of freedom) which subsumed all geographical variations and polulation density was strongly collinear with Police force. Model 10 was also not preferred as preference was given to models that have a combination of meteorological variables which can identify their impact on the number of road accidents. Among the others, model 19 and 20 had smaller VIF for the meteorological factors but their *BIC* was not better than model 15, so they were not preferred.

Model 15 was also of special interest as the coefficient of rain was found to be non-significant when GLM was used. It was observed in coming section, when GEE with AR1 error structure was used to accommodate the presence of serial correlation the coefficient of rain became significant. This led to further investigations to identify any further changes in the parameters estimates according to this model formulation.

The results of the analysis of temporal effects in section 3.6.1.2 also showed that in model 15 no substaintial systematic temporal trend remains that can be represented by further quadratic temporal terms in the model. Split sample tests also showed that paramters estimated by model 15 are reliable and consistent.

However, the Durbin-Watson test results in section 3.6.1.5 showed that serial correlation exists in data due to which GEE with AR1 error structure was preferred over the GLM as it

can account for the presence of serial correlation in the data. In the coming section the coefficients of model 15 with GEE-AR1-Negative binomial are compared informally with those of the GLM-Negative binomial in order to investigate whether any significance levels of the coefficients have changed. The following section describes the further analysis which was carried out on the results obtained from the model 15.

3.6.1.6.1 Comparison of coefficients for Dataset 3

The Stata software was used to estimate the coefficients of all variables. As the models were fitted to the same data, the estimates of corresponding parameters are not mutually independent. Due to this, no formal *T* test could be undertaken. In this section the estimated coefficients and the *t* values are compared and discussed informally.

It is found that the coefficients and the sign for some of the variables differed between the models (GLM and GEE). This has happened as the GEE-AR1 model was able to represent some of the meteorological variables through the autoregressive error term. The variables which were not related to weather (time, population density, vehicle per head of population) had not changed their signs whereas the sign of September and minimum temperature have changed. It was observed that some variation in the coefficients of month between GLM and GEE-AR1 occurred. Observing the coefficient values of GEE-AR1, it is found that the coefficient of month decreased for the winter months whereas it increased for some months of Spring and Summer (March, April and September). The coefficient of March which was not significant in GLM turned to be significant in GEE-AR1 model.

From this, it is concluded that in the time series model the partial effect of minimum temperature can be represented through the month. A few other models (models 17, 19 and 23), with various combinations of variables using a GEE-AR1 error structure, were used and coefficients of month from these models were compared with Model 15. It was found that the pattern of month of year variables remained consistent through various GEE-AR1 models: September had the same sign in all the models. This further confirmed that some of the meteorological effect is represented adequately by month in the time series models. However, once AR1 error structure is allowed, the effect of variations in rainfall over and above mean values becomes statistically significant. This will affect both police force areas that generally

have rainfall different from the national mean and times when rainfall differs from the monthly mean.

From the results of GEE model 15, when AR1 error structure was allowed it was observed that November has greatest risk of road accident per unit of distance travelled than other months whereas April has the lowest risk. The coefficient of time showed that the road accident risk per unit of distance travelled decreased at about 1 percent per annum. Population density had a positive coefficient which indicates that police forces having a higher population density tended to have greater risk of road accidents per unit of distance travelled. The geographical areas where the vehicle per head of population is high will tend to have less risk per unit of travel. Rainfall above the monthly mean is associated with more risk of having road accident per unit of distance travelled. On the other hand, increase in the mean minimum temperature is associated with less risk per unit of travel. Figure 3.6, 3.7 and Table 3.7 show the comparison of coefficients.

Figure 3.6: Comparison of coefficients of Model 15 with GLM and GEE-AR1

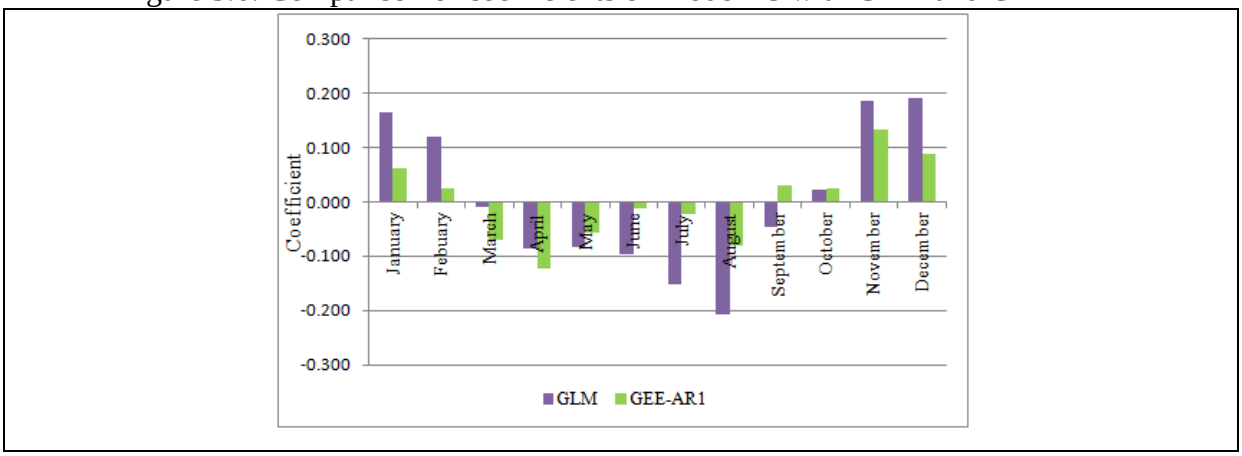


Figure 3.7: Comparison of coefficients of month by Model 15, 17, 19 and 23 (GEE-AR1-NB)

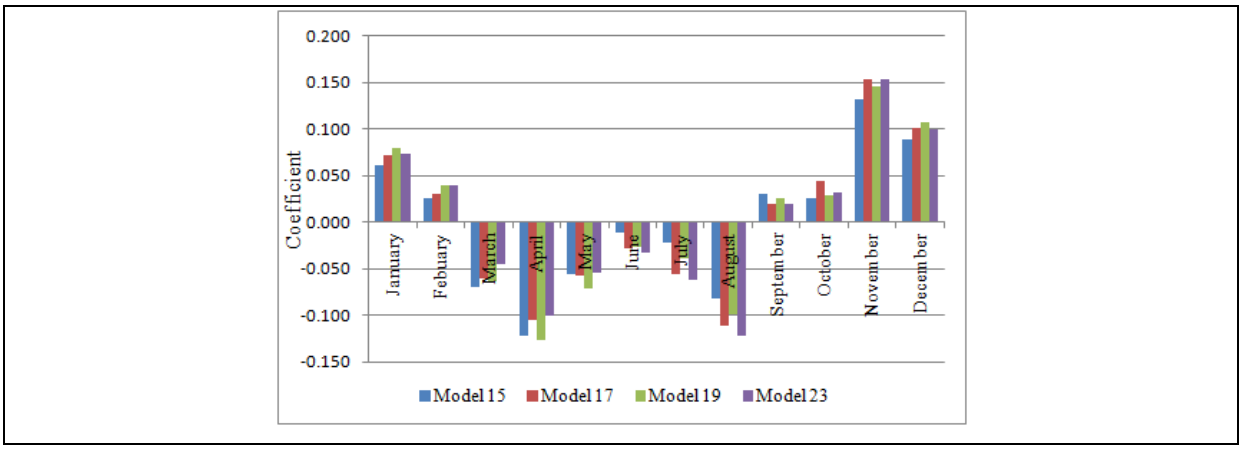


Table 3.7: Comparison of coefficient and *t* values of GEE-AR1 and GLM-Model 15-NB for coefficient validation (Dataset 3)

Variables	Comparison of models			
	Model 15-GEE-NB(AR1)		Model 15-GLM-NB	
	Coefficient	*t* value	Coefficient	*t* value
January	0.061	5.57	0.165	12.44
February	0.025	2.31	0.120	9.14
March	-0.069	-7.31	*-0.010*	*-0.85*
April	-0.122	-14.17	-0.086	-8.20
May	-0.056	-6.80	-0.084	-8.26
June	*-0.011*	*-1.12*	-0.096	-7.85
July	*-0.022*	*-1.77*	-0.152	-10.23
August	-0.081	-6.61	-0.208	-14.07
September	0.031	3.09	-0.047	-3.90
October	0.025	3.06	0.021	2.10
November	0.131	14.82	0.186	17.13
December	0.088	8.12	0.190	14.53
Time	-0.001	-6.77	-0.001	-15.69
Pop density	0.0002	9.05	0.0002	18.74
Veh/Person	-1.290	-20.44	-1.310	-41.00
Min temp	-0.008	-4.79	0.014	7.27
Rain	0.001	10.54	*7.4E-05*	*0.98*
Constant	-13.908	-454.60	-13.993	-817.19

Italic shows that these variables are not significant at 5 percent level

3.6.1.6.2 Comparison of the number of road accidents observed and estimated for each month, Standardized deviance residuals and cumulative percentage graphs

Graphs of road accidents observed and estimated for each month within a police force area showed good agreement. However, from the graph and subsequent analysis of the results in Table 3.8 it was observed that the model was not reliable when the road accidents for a month were either less than 100 or greater than 800. The cumulative proportion graph in Figure 3.8 also confirms this. In the whole dataset there were 130 observations when road accidents were observed to be lower than 100. The estimated values gave only 44 such months. In the same way there were 45 observations when number of road accidents for a month was higher than 800 whereas the estimated values gave only 24 such months. The summary of the number of road accidents observed and estimated for each month is shown in Table 3.8.

The standardized deviance residual graph in Figure 3.8 showed that Grampian had the most negative standardized deviance residuals. It was observed that out of the 100 most negative standardized deviance residuals (SDR), 67 belonged to Grampian. The reason for this might be that Grampian had the lowest number of road accidents and the model estimated slightly higher values for this police force. The highest negative SDR was -3.96 which occurred in March 2000 for the Grampian police force, where observed road accidents were 68 while the model estimated it to be 153. Another outlier in July 2004 was from the same police force, where the numbers of observed and estimated road accidents were 65 and 145 respectively. There were a few outliers with the highest positive value which mostly belonged to the months of December and January. The highest positive outlier was for Cambridgeshire police force in the month of December 2002 where 341 road accidents were observed compared to an estimated 175 accidents. Generally, it was observed that SDR lies between -4 and +4.

Figure 3.8: Number of monthly road accidents observed and estimated, Standardized deviance residual graphs (Dataset 3)

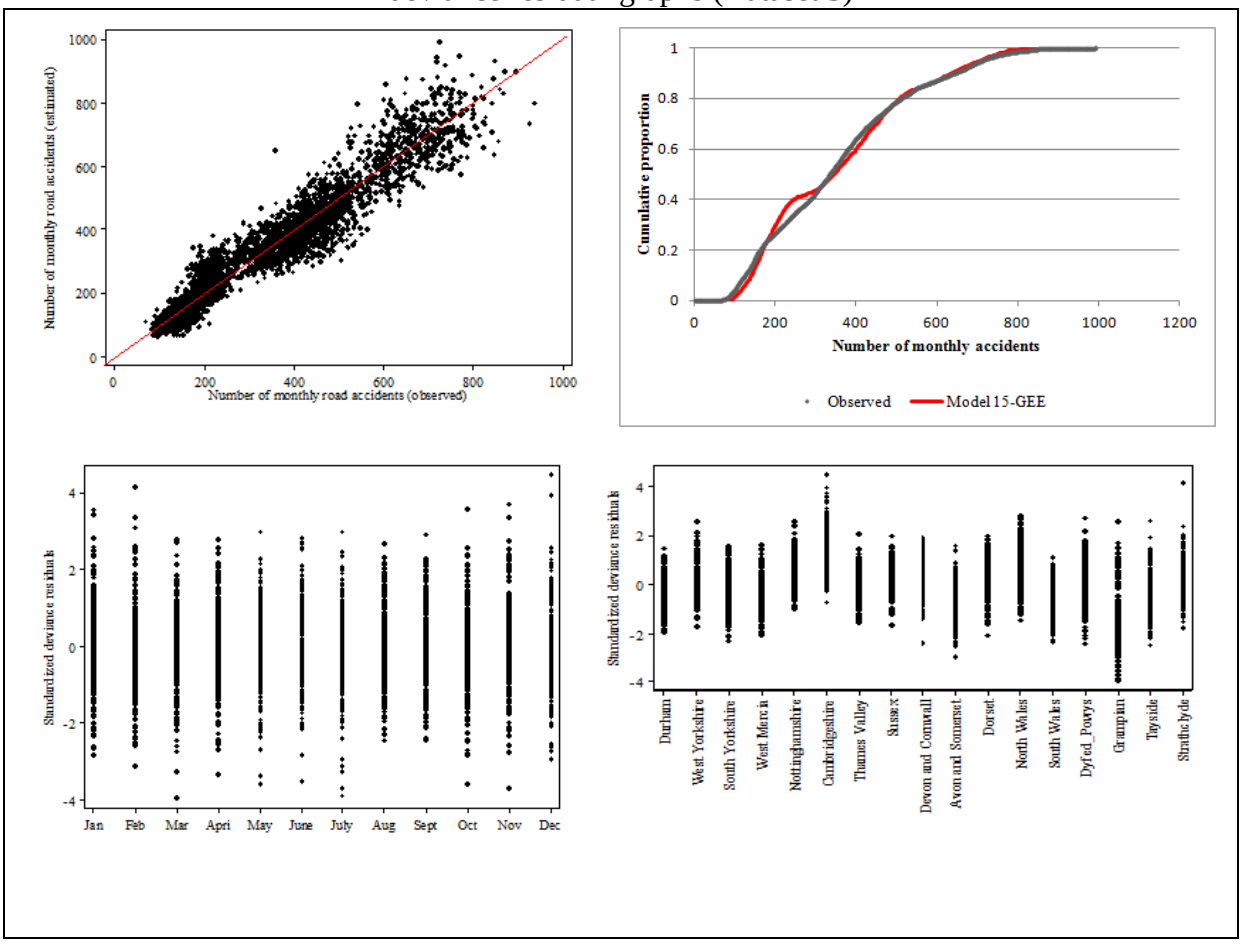


Table 3.8: Summary of road accidents observed and estimated (Dataset 3)

Group	Number of observation in the group	
Numerical Range of monthly road accidents (Groups)	Number of monthly road accidents observed	Number of monthly road accidents estimated
0 to 100	130	44
100 to 200	686	866
200 to 300	484	426
300 to 400	660	497
400 to 500	450	575
500 to 600	258	249
600 to 700	214	239
700 to 800	133	140
800 to 900	39	22
900 to 1000	6	2

3.6.1.6.3 Final model checking graphs:

Some graphs were plotted in Figure 3.9 to check visually if any problems existed in model 15 with the GEE-AR1 error structure. The first graph shows the deviance residuals plotted against the fitted values. It is observed that plot does show some trend as well as a substantial variation in the density of observations over the range of fitted values. The greatest negative residuals occur when the number of road accidents estimated is about 150 to 200, which correspond to observations from the Grampian police force. The greatest positive residuals were found for estimated road accidents ranging in number from 180 to 250, which were found mostly to come from the Cambridgeshire police force: other variations appear to stem from police force areas. In order to investigate the nature and strength of this variation in the deviance residuals, the averages of the absolute values of these residuals were calculated in bands of 50 of the estimated values: the results of this are plotted in Figure 3.10. This shows a generally decreasing trend in magnitude of deviance residuals with increasing fitted value.

As a result of higher deviance residuals for police forces of Cambridgeshire and Grampian as shown in deviance residuals and fitted values graph further investigation was carried out. It was observed as shown in Figure 3.3 that Cambridgeshire had lowest amount of rainfall whereas Grampian had lowest minimum temperature among all the police forces considered in this study. In light of this, a test was carried out to investigate the effect of adding the two

further explanatory variables of rain in Cambridgeshire and minimum monthly temperature in Grampian to model 15 with GEE-AR1 error structure. The deviance residuals estimated from this refined model do not show any substantial improvement relative to the plot of deviance residuals against predicted values which is shown in Appendix A3.3, hence this refinement was not considered further.

The second graph considered was the normal quantile plot of standardized deviance residuals. From the graph it is observed that the quantile plot follow the straight line closely, supporting the assumption of normality of the residuals. Some minor deviations are observed especially at the ends, which suggest the data distribution had a long tail at each end. The scale location plot also showed that deviance is slightly decreasing with increase in fitted values, though a substantial variation in the density of observations over the range of fitted values was also observed.

In the last graph Cook's distance plot shows most of the observations that had higher peak relates to the months of November, December and January. Noticeably, the highest peak was observed which represents the observations from the Cambridgeshire police force. However, the Cook's distance for these observations was less than the critical value of 1 that would cause concern.

The graphs shown in Figure 3.9 and 3.10 suggest presence of heteroscedasticity in the residuals. In order to confirm this, Park and Glejser tests were carried out. The test results shown in Appendix A3.4 verify that heteroscedasticity is present in the residuals. Due to this an adjustment to the standard errors of coefficients was made using the White's procedure as implemented in STATA. However, we note that the hierarchical generalized linear model (HGLM) introduced and used in Chapter 5 allows to model variations in dispersion.

In Table 3.9 the results of model 15 using GEE-AR1 are compared after adjusting the standard errors due to the presence of heteroscedasticity. The results show that *t* values of all the variables have decreased typically by a factor of 2 except October and rain which have increased slightly. The coefficient of February turned to be non-significant after implementing the corrections. This suggests that if the presence of heteroscedasticity is not accounted the coefficients will not be efficient but they will still be unbiased and consistent.

Figure 3.9: Diagnostic plots for model 15 (Dataset 3)

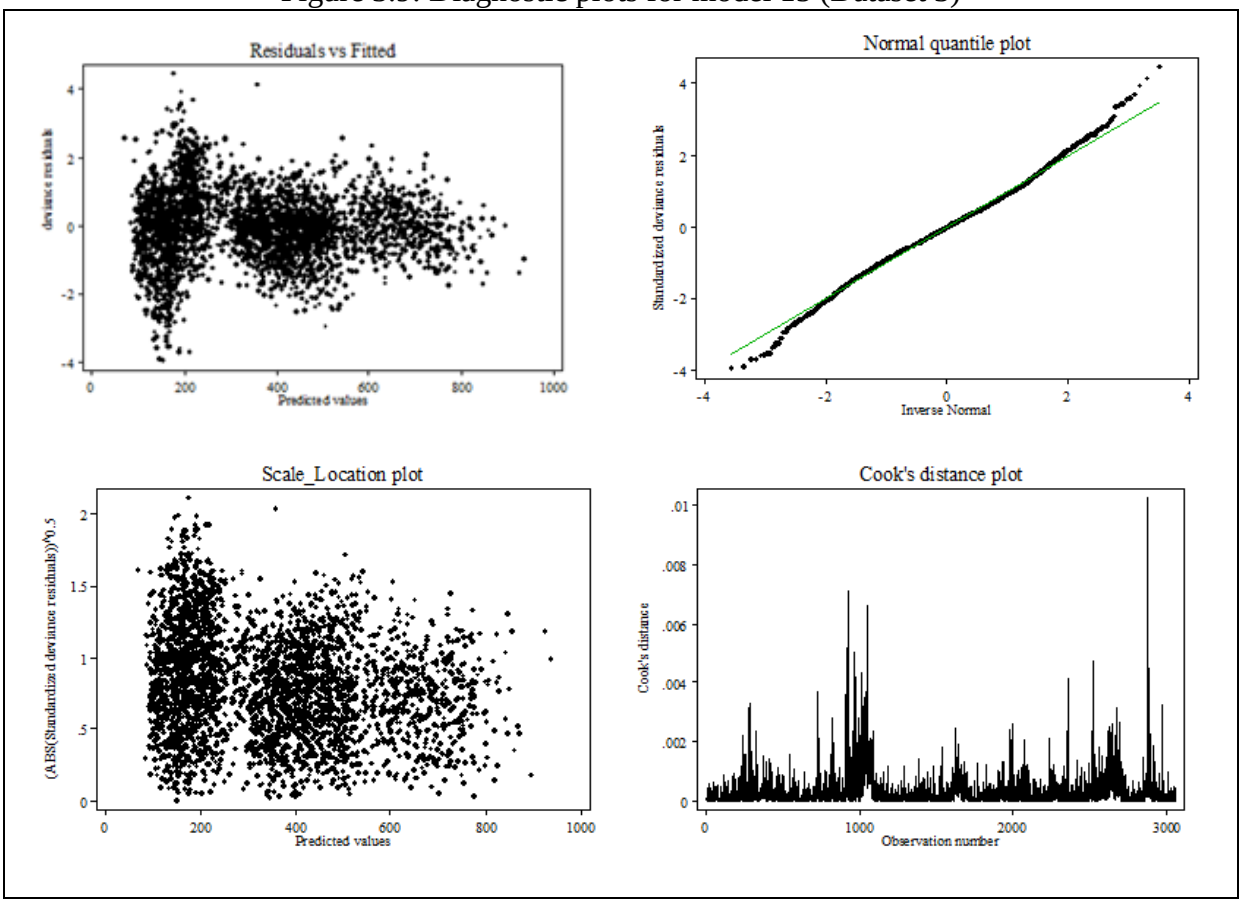


Figure 3.10: Average of the absolute value of deviance residuals and estimated values in bands-Dataset 3

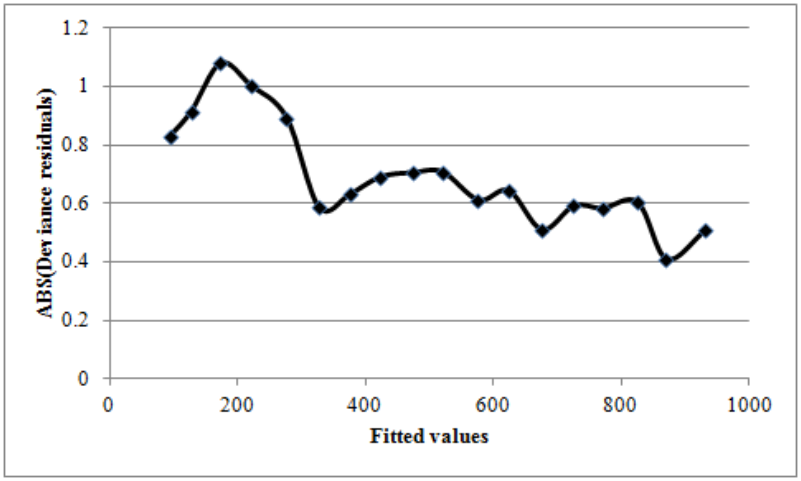


Table 3.9: Comparison of coefficient and *t* values of GEE-AR1 Model 15-NB after using correction for the presence of heteroscedasticity

Variables	Comparison of results of model 15-GEE-AR1			
	Before applying an corrections		White's Robust Standard Errors	
	Coefficient	*t* value	Coefficient	*t* value
January	0.061	5.57	0.061	3.11
February	0.025	2.31	*0.025*	*1.17*
March	-0.069	-7.31	-0.069	-5.98
April	-0.122	-14.17	-0.122	-14.83
May	-0.056	-6.80	-0.056	-5.64
June	*-0.011*	*-1.12*	*-0.011*	*-0.75*
July	*-0.022*	*-1.77*	*-0.022*	*-0.78*
August	-0.081	-6.61	-0.081	-2.74
September	0.031	3.09	0.031	2.41
October	0.025	3.06	0.025	3.27
November	0.131	14.82	0.131	8.26
December	0.088	8.12	0.088	3.84
Time	-0.001	-6.77	-0.001	-3.02
Pop density	0.0002	9.05	0.0002	3.15
Veh/Person	-1.290	-20.44	-1.290	-11.61
Min temp	-0.008	-4.79	-0.008	-2.04
Rain	0.001	10.54	0.001	11.33
Constant	-13.908	-454.60	-13.908	-204.38

Italic shows that these variables are not significant at 5 percent level.

3.7 CONCLUSION

The purpose of this part of the study was to assess the impact of meteorological variables on the risk of road accidents per unit of travel. A specific objective was to determine whether meteorological variables contribute to the variability in the number of road accidents among the months. This was undertaken using road accident data for each police force area during each month. The adjusted distance travelled in the month for each police force was used as the offset which accounted for the variations in distance travelled during the months of year. As a result of this, the linear predictor in this model can be interpreted in terms of an estimate

of the risk of road accidents in a police force area during each month per vehicle-kilometre of distance travelled there.

The results showed that serial correlation exits in the data due to which the Generalized Estimation Equation (GEE) having autoregressive order 1 (AR1) with negative binomial was preferred to the Generalized Linear Model (GLM). In this particular case, it was observed that coefficients for the variable of Month estimated by GEE-AR1 were substantially different from the coefficients estimated by GLM. This change happened as in GEE-AR1 model it had been represented through the coefficients of the month. Observing the coefficient values of GEE-AR1 it was found that coefficient of month reduced in the winter months while it increased for some months of Spring and Summer (March, April and September). The coefficient of rainfall was also found to be statistically significant in the GEE-AR1 model. The amount of rain is associated with greater risk of road accident per unit of distance travel whereas the increase in the minimum temperature is associated with less risk per unit of travel.

It was also found that November is associated with greater risk of road accidents per unit of distance travelled than all other months of year. April had lowest risk after allowing for the meteorological effects. This finding differs from that in chapter 2 in which travel during August was reckoned to have less risk than April: this difference arises through allowance for meteorological effects. The coefficient of time is negative showing that road accident risk per unit of travel is becoming progressively less risky. Circumstantial variables that characterise the police force showed that higher population density resulted in greater accident risk and the police force areas having more vehicles per head of population had lower risk per vehicle-kilometre of travel than other police forces.

It was generally observed that inclusion of a small number of meteorological variables can improve the goodness of fit of a model. The effects of the local climate should therefore be considered before designing any systematic safety plans for a region.

4. MODELLING THE NUMBER OF VEHICLES INVOLVED IN ROAD ACCIDENTS

4.1 INTRODUCTION

Various safety improvement programmes are designed by the planning and development agencies to reduce both the number of road traffic accidents and the severity of those that do occur. The numbers of road accidents are estimated by using road accident prediction models. These models relate the expected number of road accidents to some available explanatory variables. Based on modelling results, appropriate road safety initiatives can be proposed to improve road safety. If the initiatives are inappropriate, this can result in reduced road safety and waste of resources. The several techniques available for estimating the number of road accidents have been described in detail in Chapter 2 and are summarised below.

In earlier research (Andrecscu and Frost, 1998; Bester, 2001) the relationship between road accidents and other variables was found by using a conventional multiple regression technique. As noted earlier, this approach lacks the distributional properties that are appropriate to adequately describe random, discrete, and non-negative events such as traffic accidents. Various studies including Miaou and Lum (1993) and Miaou (1994) have shown that test statistics derived from these models are questionable because they do not necessarily use the appropriate distributions. Maycock and Hall (1984), and Maher and Summersgill (1996) have shown that variance of count data is found to be higher than the mean; the extra variation is known as over-dispersion. When using Poisson regression in the presence of over-dispersion, model parameter estimates will still be close to their true values, but their variance of estimation tends to be under-estimated and the significance levels of estimated coefficients will therefore be overstated. This has been addressed by Hadi et al (1995) and Anis (1996) who have shown significant advances in describing the discrete traffic accident count data by producing more accurate and reliable models through the use of generalized linear models with Poisson and negative binomial distributions. In order to address the issue of over-dispersion, Abdel-Aty and Radwan (2000), Guevara et al (2004), McCarthy (2005) used the negative binomial distribution which allows variance to exceed the mean.

Another important issue for the time-series of road accident data arises through the presence of serial correlation. In the presence of serial correlation, efficiency of the parameter

estimates comes into question. Lord and Persaud (2000) used the generalized estimation equation (GEE) methodology which has the additional capability to accommodate temporal correlation in the data. Wang and Abdel-Atey (2006) used GEE to accommodate serial correlation in data for modelling road accidents at different intersections. Memon (2008) used GEE with AR1 error structure for modelling the number of vehicles involved in road accidents in Great Britain. Ulfarsson and Shankar (2003) used a negative multinomial (NM) model to account for the panel structure of the data that arises from repeated observations at each set of sites.

From this literature review we conclude that the generalized linear model (GLM) with Poisson error structure and logarithmic link function goes some way to addressing the requirements of modelling the numbers of vehicles involved in road accidents. However, this approach does not accommodate the over-dispersion that is encountered in these counts, and this leads to overstatement of accuracy of parameter estimates. Furthermore, this model structure does not accommodate the serial correlation that is also encountered. Use of the negative binomial error structure can accommodate over-dispersion, and use of AR1 time series error structure can accommodate serial correlation. Together, these extensions to the statistical model will lead to improved estimates of parameters and their accuracy. These features are provided by the GEE model formulation.

The present research has the following objectives;

- To compare the results of generalized linear models and generalized estimation equations in order to develop road accident prediction models which can accurately estimate the number of vehicles involved in road accidents on each day disaggregated by road class and vehicle class in Great Britain based on the national accident dataset of STATS 19.

- To identify the relationship between the numbers of vehicles involved in road accidents on each day and other variables such as road class, vehicle class, day of the week, month, time and various holidays.

- To estimate the risk of involvement in a road accident per unit of travel for different road and vehicle combinations.

The investigation presented in this chapter focuses on the combined use of road accident and vehicle information from STATS 19 data along with traffic flow data. The presence of serial correlation due to the natural order of observations will also be tested and it will be observed whether this affects estimates of the parameters of the models and the associated test statistics. Models were initially developed using GLM with a negative binomial regression. For the preferred model, GEE-AR1 is used to accommodate serial correlation and the results are compared with GLM.

This study identifies a suitable technique to model the number of vehicles involved in road accident datasets using GEE-AR1. The estimated risk values of being involved in a road accident per unit of exposure for all road and vehicle combinations can be used to highlight those combinations that need most attention. The results of this research will help various planning and emergency rescue agencies to develop road safety intervention programmes for targeted road and vehicle combinations and to identify significant variables in an appropriate way. This will also enable agencies to allocate the resources and focus on particular road user groups in an efficient way by anticipating how many vehicles are likely to be involved in road accidents on any day by road class throughout the study area. The results obtained from this study may also help to promote education and safer use of road and vehicle combinations.

This chapter is organized as follows. Section 4.2 describes the data used for this study, which is analysed briefly in Section 4.3. Section 4.4 presents the process of model development and basic structure of the model. Section 4.5 shows the model selection process, results of developed models, goodness of fit and model checks. Section 4.6 presents the resulting estimated risk per unit of travel for various vehicle classes. Finally some concluding remarks are given in section 4.7.

4.2 DATA USED

The STATS 19 road accident and vehicle data, and the traffic flow data that are used for this study are described below:

4.2.1 Combined road accident and vehicle data (STATS 19 data)

In this part of the study a new dataset, denoted as Dataset 4, was developed which had number of vehicles involved in road accidents for each day instead of the earlier dataset used in Chapter 2 which had the number of road accidents for each day, for the following reasons:

- Additional information regarding the road and vehicle class combination was to be explored through this modelling. Information relating to vehicle and road class is not available in the accident section of STATS 19 which was the source of information in the earlier dataset. In the present case, data from the accident and vehicle section of STATS 19 data were combined and a new dataset was formed to represent the number of vehicles involved in road accidents on each day by road and vehicle class rather than road accidents on each day.
- No suitable corresponding traffic flow information was available for Dataset 2 as that data related to the number of road accidents occurring on each day for police forces of Great Britain.

The road accident statistics in Great Britain are compiled by the police. All road accidents involving human death or personal injury occurring on the highway are required to be notified to the police within 30 days of occurrence. For each such road accident, police authorities complete a STATS 19 form which provides details of road accident circumstances, information on each vehicle involved, and information of each person injured in the road accident. This whole dataset is maintained by the Department for Transport (DfT). In the present chapter, the five years' road accident data from 2001 to 2005 was used for modelling the involvement of vehicle classes in accidents on different road classes.

Before combining the information from the accident and vehicle sections of the STATS 19 data and the traffic flow data which was obtained from the DfT, the distinct road

classifications were reconciled. In order to make joint use of the two different sources of information, roads were reclassified in STATS 19 data by using the speed limit information. In STATS 19 data, roads are classified as: motorway, A, B, C, and unclassified whereas available traffic flow data from the DfT are classified as: motorways, rural A, urban A, rural minor, and urban minor roads. Thus MS Access queries were used to reclassify the roads as shown in Table 4.1. It was also found that the vehicle classification of STATS 19 does not match that is used in the traffic flow data. Due to these limitations, vehicles classes were also reclassified as shown in Table 4.2. The vehicle classes of minibus, other motor vehicles, other non-motor vehicles, ridden horse, agricultural vehicle and tram were excluded from the dataset for this study because of the unavailability of traffic flow data and their involvement in only a few road accidents.

After reclassification, extensive work was done to combine the accident and vehicle sections of STATS 19 data for each year. It should be noted that for each road accident there were one or more vehicles involved. These two sections of road accident data were joined by using the accident reference number. For this process MS Access and SPSS were used. These Access files were exported to SPSS to develop a new dataset which consisted of the information about all vehicles involved in road accidents from 1st January 2001 to 31st December 2005. SPSS cross-tabulations were used to extract the information for the number of vehicles involved in accidents for each day by road class and vehicle class. All this was done to bring the road class and vehicle class variables into the new dataset as the accident section has no information about the road class and vehicle class, and the vehicle section on its own could not identify when and where the road accident happened. After combining them the information of road class, vehicle class, day, month, and year were available in a single dataset. A total of 24 different combinations were used. The dataset contains five years' information of vehicles' involvement in road accidents. It had total of 43,824 observations for all 24 different groups. Each group represents a different vehicle class and road class, and has 1,826 observations each representing the number of vehicles involved in road accidents on each day by road type and vehicle class from 2001 to 2005. The group involving pedal cycles on motorways was excluded from the dataset because pedal cycles are not allowed on motorways and so are rarely if ever involved in road accidents on them.

Table 4.1: Criteria for rearranging road classification

S.No	Roads reclassified		
	New classification	Criteria	
		STATS 19 data classification	Speed limit (mph)
1	Motorway	Motorway	-
2	Rural A	A(M) or A	> 40
3	Urban A	A(M) or A	≤ 40
4	Rural Minor	B or C or Unclassified	> 40
5	Urban Minor	B or C or Unclassified	≤ 40

Source of data: Department for Transport (2011)

Table 4.2: Vehicles classes used for the study

S.N	Vehicles classified in STATS 19	New classification	S.N	Vehicles classified in STATS 19	New classification
1	Pedal Cycle	Pedal cycle	9*	Other motor vehicles	
2	Moped		10*	Other non-motor vehicles	
3	Motorcycle 125 cc		11*	Ridden horse	
4	Motorcycle > 125 cc	Motor cycle	12*	Agricultural vehicle (in diggers etc)	
5	Taxi		13*	Tram	
6	Car	Car	14	Goods 3.5 tonnes mgw or under	
7*	Mini bus (8-16 passenger seats)		15	Goods over 3.5 and under 7.5 t	Goods Vehicles
8	Bus or coach (17 or more passenger seats)	Bus	16	Goods 7.5 tonnes mgw and over	

Source of data: Department for Transport (2011)
** These vehicle classes were not included in the study*

4.2.2 Traffic flow data

Traffic flow data is obtained from the DfT which estimates the flows from the information obtained from traffic counts that are conducted at different types of road. Traffic counts are carried out manually and automatically as described below:

4.2.2.1. Manual counts: According to the DfT (2005), manual counts operate differently for major and minor roads. Roads classified as major are: motorways, trunk roads, and principal roads with the latter two divided into urban and rural roads. Roads classified as minor are the

three main classes of B, C, and U (unclassified) roads and each is subdivided into urban and rural, resulting in a total of six classes.

a. Manual counting for major roads: For major roads (motorways and A-roads) the traffic on every link is assessed regularly. Traffic counts are done at a random point on most of the links at regular intervals, once every three years in England and Wales, and once every six years in Scotland. About 5,100 major road sites were counted in 2005 (DfT, 2005). Additional information about the characteristics of each link such as its length and road width at the location of the count is also gathered. Trained enumerators count vehicles from 7 am to 7 pm. All counts take place on weekdays, but not on or near to Public holidays or school holidays. The counting is also confined to neutral weeks to minimise the effects of seasonal factors; these neutral weeks are mostly in the months of March, April, May, June, September, and October. Some major links are unsafe or too short to be worth counting in the usual manner. In these cases traffic estimates are made from the judicious use of flow data on adjacent links. These are called derived links. Some links are treated as a dependent link and defined as ending at the local authority boundary. In these cases it is assumed that the flow is the same along the entire link, so a count in one local authority can be used a proxy for the flow on the dependent link. In 2003 there were 15,500 normal links, 1,200 derived links and 1,000 dependent links.

b. Manual counting for minor roads: Minor road traffic estimates are made by grouping minor roads into six road classes. The average flow on each of these road types is estimated by carrying out the several counts along them. A sample of about 4,500 sites across Great Britain is visited each year on neutral weeks. These same sites are counted each year. Apart from this, 200 counts per year are carried out in non-neutral weeks and on weekends which are known as summer-winter counts. These counts provide extra information about two-wheeled traffic throughout the year, as pedal cycles and motorcycles are not always accurately identified by automatic counters.

4.2.2.2. Automatic counts: There are 190 sites in Great Britain outside London where traffic is monitored continuously using automatic sensors which classify the traffic into vehicle type. The automatic counting equipment recognises 22 different types of vehicles which are then combined to provide estimates for the 11 vehicles types used by the DfT. These counters are not fully accurate as they cannot correctly classify traffic moving at 5 mph or less. The

automatic counters in London are slightly different to those outside London. In London, there are 54 counters and they are volumetric classifiers as they only distinguish between short (up to 5.2 metres) and long (greater than 5.2 metres) vehicles. These counters need 24-hour manual counts every three months to provide estimates of the breakdown of traffic by vehicle type in each hour of the day.

4.2.2.3. Annual average daily flows (AADF): The data for all manual counts in neutral months are combined with information from automatic counters on similar roads to provide an estimate of the AADF at that site. This is normally done by multiplying the raw count data by factors derived from automatic counts in that same year. There are a large number of correction factors, for each vehicle type, day of counting, and various other groups. As these counts are done in neutral weeks, the expansion factors used do not vary too much from year to year except when bad weather has restricted traffic during the winter months.

4.2.2.4. Estimating annual traffic estimates from AADFs: Different procedures are applied for major and minor roads in converting AADF data to traffic estimates. For every major road link its AADF is multiplied by its length and the number of days in the year to get the value in million kilometres per year. As every major road link is counted, so a summation of all the links will lead to annual traffic estimates. For each minor road class in each local authority area an AADF is estimated based on a sample of traffic counts. These AADFs are then multiplied by the total road length for the relevant minor road category to give an estimate of traffic in vehicle-km for that road category.

4.3 DATA ANALYSIS

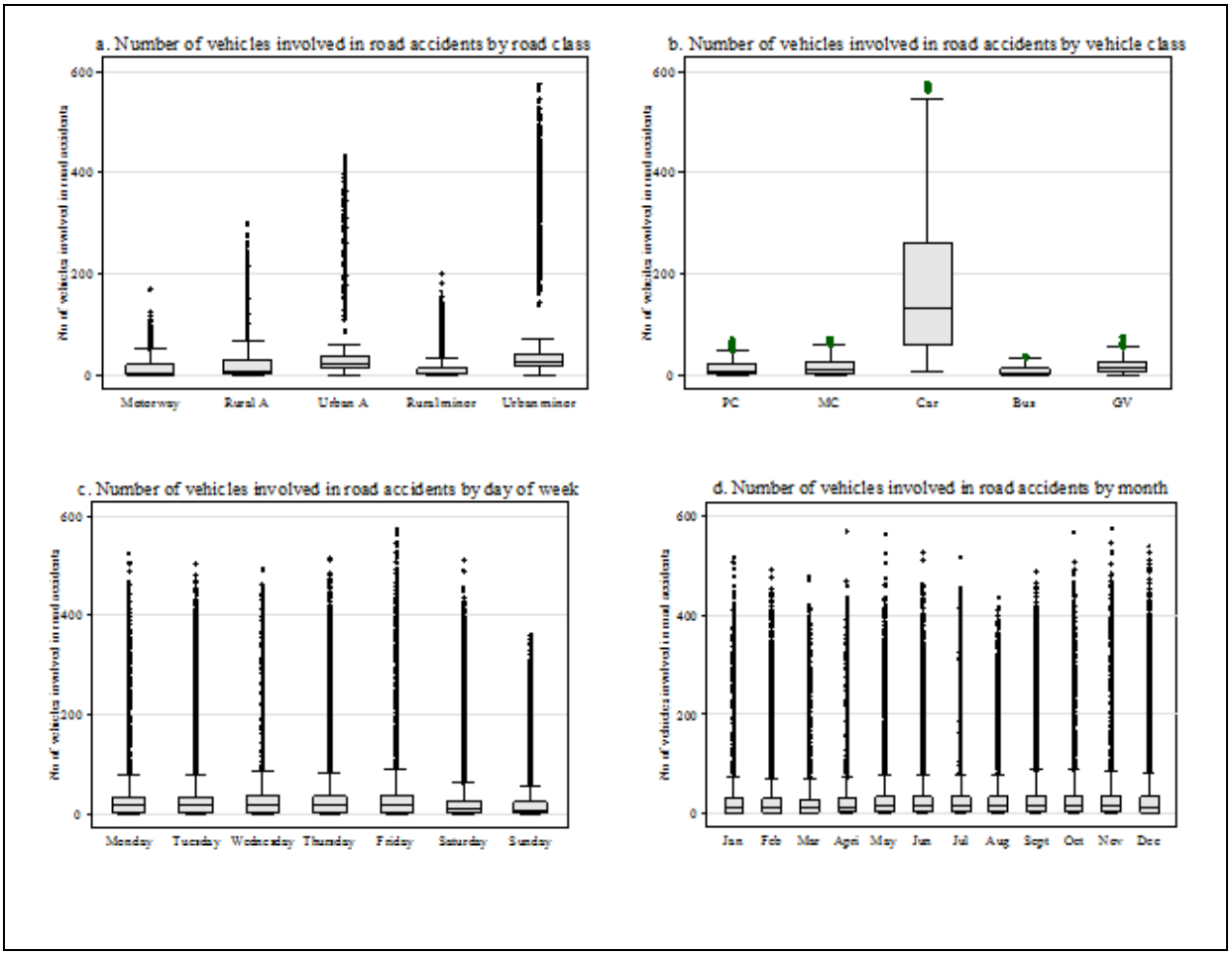
STATS 19 data, road length data, and traffic flow data used for this study are analysed as follows:

4.3.1 Analysis of STATS 19 data

The combined STATS 19 data of accident and vehicle section for 2001 to 2005, which represents the number of vehicles involved in road accidents occurring on each day, was analysed by using box plots produced by Stata software which are shown in Figure 4.1 and explained as follows:

It was observed that more vehicles were involved in road accidents on urban roads. A wide disparity exists among road classes in terms of highest number of vehicles involved in road accidents. The rural minor roads had a lower interquartile range which indicates less variability in terms of the number of vehicles involved in road accidents. Figure 4.1 also shows the prevalence of cars involved in road accidents. The median for cars involved in road accidents was at least nine times higher than the median of all other classes of vehicles. The interquartile range for cars indicates that the level of involvement of cars in road accident on different roads may also vary a lot. A slight difference is observed between weekdays and weekends. Day-to-day variation in terms of numbers of vehicles involved in road accidents was not so great but Sunday had a lower median than other days. December and January had the lowest median among all the months. It was also observed that the initial four months of the year had a lower median than later months, except December, which might be due to seasonal differences.

Figure 4.1: Box plots of STATS 19 data (Dataset 4: 2001 to 2005)



Source of data: Department for Transport (2011)

4.3.2 Analysis of road length data

The road length data of all road classes for 2001 to 2005 used in this study was obtained from the DfT. The figures showing the yearly length of all road classes are shown in Table 4.3 and were incorporated into Dataset 4. It was found that:

- Motorways had the lowest proportion of roads equalling almost 1 percent of the total road length. This proportion was in range of 0.88 to 0.91 percent for all the years.
- A-class roads were 12 percent of the total road length. Rural A roads were three times longer than urban A roads. The length of rural A and urban A roads were about 9 and 2.8 percent respectively out of the total road length.
- Minor roads were 87 percent of the total road length. Rural minor roads constituted about 54 percent whereas urban minor roads were 34 percent of the total road length.

As shown in Table 4.3, the road lengths were similar over the years but it was found that total road length for 2004 and 2005 was less than for the initial three years. According to the DfT report Transport Statistics for Great Britain, 2007 this is mainly due to amendments made to road lengths in Scotland as some of the private roads maintained by the Forestry Commission were earlier recorded as public roads.

Table 4.3: Road length of various road classes (2001-2005)

Road class	**Year**				
	2001	**2002**	**2003**	**2004**	**2005**
Motorway	3,476	3,478	3,478	3,524	3,520
Rural A	35,522	35,532	35,525	35,530	35,550
Urban A	11,132	11,141	11,127	11,138	11,107
Rural Minor	210,037	210,343	210,656	207,565	207,646
Urban Minor	130,802	131,169	131,556	129,917	130,186

Source of data: Department for Transport (2011)
**road length in kilometres*

4.3.3 Analysis of traffic flow data

Traffic flow data was obtained from the DfT for different road and vehicle combinations which were jointly used with STATS 19 data. For the purpose of understanding, data was aggregated in this section to determine the share of each road and vehicle class in the total yearly distance travelled. The aggregated results showed that over 90 percent of total yearly vehicle kilometres are travelled by car or taxi. The proportion of distance travelled using pedal cycles, motorcycles, and buses, with slight yearly variations, was about 1 percent each out of the total yearly distance travelled. Higher distances were travelled by goods vehicles which constituted about 7 percent of the total vehicle kilometres travelled. On the other hand, road class aggregation of data revealed that although motorways were 1 percent of the total road length in Great Britain, 19 percent of the total distance was travelled on these roads. Rural A road which constituted 9 percent of the total road length carried 28 percent of total traffic. It was also found that although minor roads (either rural or urban) constituted 87 per cent of the total network of Great Britain, only 37 percent of the total yearly distance was travelled on them. Table 4.4 gives the percentage of distance travelled for each road class and vehicle combination from the total yearly distance travelled for the years 2001 to 2005. This table shows that:

- Car and taxis were the dominant form of traffic on motorways and on all roads, with shares ranging from 85 percent on the motorway and up to 92 percent on each of urban roads and rural minor roads.
- All vehicles travelled more on urban A roads than on rural A roads and motorways, except goods vehicles. The proportion of distance travelled by goods vehicles on urban A roads further reduced to about 4 percent.
- On rural minor roads, pedal cycles and motorcycles travelled slightly more than on A roads. Goods vehicles travel about 3 percent of the total distance. The proportion of distance travelled by cars stayed nearly same as on urban minor roads.
- Goods vehicles constituted the second largest form of traffic on all roads except urban minor roads. The proportion of traffic constituted by goods vehicles decreased from 14 percent on motorways to 2 percent on urban minor roads.

Table 4.4: Percentage of the distance travelled by road class and vehicle class, 2001 - 2005

Code	*Vehicle class*	Year				
		2001	**2002**	**2003**	**2004**	**2005**
		Motorway				
1	PC	-	-	-	-	-
2	MC	0.47	0.47	0.50	0.46	0.47
3	Cars & Taxis	84.39	84.93	85.0	84.76	84.99
4	Bus	0.70	0.57	0.56	0.54	0.54
5	GV	14.44	14.03	13.95	14.24	14.01
		Rural A				
6	PC	0.17	0.16	0.11	0.10	0.11
7	MC	1.04	1.03	1.11	1.04	0.99
8	Cars & Taxis	89.29	89.72	89.83	89.90	90.0
9	Bus	0.78	0.78	0.78	0.70	0.71
10	GV	8.72	8.31	8.17	8.25	8.19
		Urban A				
11	PC	0.72	0.68	0.84	0.75	0.75
12	MC	1.29	1.38	1.55	1.34	1.34
13	Cars & Taxis	92.23	92.32	91.92	92.18	92.20
14	Bus	1.73	1.69	1.63	1.56	1.60
15	GV	4.03	3.93	4.04	4.17	4.10
		Rural minor				
16	PC	1.16	1.44	1.54	1.40	1.53
17	MC	1.62	1.53	1.45	1.37	1.57
18	Cars & Taxis	92.81	92.19	92.18	92.62	92.29
19	Bus	0.93	1.42	1.38	1.24	1.09
20	GV	3.48	3.42	3.46	3.37	3.52
		Urban minor				
21	PC	2.48	2.85	2.84	2.38	2.85
22	MC	1.33	1.53	1.84	1.69	1.90
23	Cars & Taxis	92.46	92.12	91.43	92.02	91.46
24	Bus	1.51	1.72	1.97	2.02	2.02
25	GV	2.22	1.77	1.91	1.89	1.78

Source of data: Department for Transport (2011)
The numbers shown are in percentage

4..4 CORRECTIONS APPLIED TO TRAFFIC FLOW DATA TO ADJUST FOR DAILY AND MONTHLY VARIATIONS

As the traffic flow data varies by the day of the week and month of the year, this variation was taken in account to some extent by using day of week and monthly correction factors to adjust the traffic flow data for each day. These correction factors for each year from 2001 to 2005 were obtained from DfT and were derived from continuous automatic counts conducted at a small number of fixed sites on major and minor roads as explained in section 4.2.2. Slight adjustments as explained below were made to make these correction factors compatible with our dataset.

- **Road classification:** The correction factors were available for four categories of roads, these being: motorways, all rural major and minor roads, all urban major and minor roads, and all roads. In this case instead of a single correction factor for all roads, separate ones were used for rural roads and urban roads. This adjustment was based on the assumption that the traffic flow on major and minor roads varied in a similar way by day of the week and month of the year which was near to the ideal situation when correction factors for all the five classes of road (Motorway, Rural A, Urban A, Rural minor and Urban minor) could have been used.
- **Vehicle classification:** The correction factors for cars and taxis, goods vehicles, and all motor vehicles were available. In this case the correction factors for all motor vehicles were applied with the assumption that traffic flow in each vehicle class varies in the same way on different roads. This assumption seems far from the ideal of using separate correction factors for each of pedal cycles, motorcycles, cars and taxis, buses, and goods vehicles.

Due to the limitations on availability of day of week and month correction factors, factors for all motor vehicles on motorways, all rural major and minor roads, and all urban major and minor roads were used to adjust for variation in the number of vehicles involved in road accidents. The correction factors for the year 2005 are shown in Table 4.5, which shows that on Fridays a higher distance was travelled on all roads whereas on Sundays the lowest distance was travelled. In August the greater distance was travelled on motorways and all rural major and minor roads whereas a greater distance was travelled on all urban major and

minor roads in March, April, and November. In December, January, February, March usually less distance was travelled on all roads in comparison to other months.

Table 4.5: Daily traffic flows by day of the week and month of the year (2005)[1]

Index: Average daily traffic = 100

Day of week	**Road classes**		
	Motorways	**All rural major and minor roads**	**All urban major and minor roads**
	All motor vehicles	All motor vehicles	All motor vehicles
Monday	104	103	102
Tuesday	104	103	105
Wednesday	105	104	107
Thursday	108	107	108
Friday	114	114	110
Saturday	82	90	92
Sunday	83	79	75
Month of year			
January	91	87	96
February	94	91	97
March	98	97	102
April[2]	101	101	102
May	100	103	101
June	103	105	101
July	105	107	101
August	107	110	98
September	105	106	101
October	103	102	101
November	100	99	102
December[3]	93	92	97

Source of data: Department for Transport (2011)

1. Indices are based on average daily traffic and are not affected by the varying number of days in each month.
2. Figures are affected by Easter
3. Figures are affected by Christmas

4.5 MODEL DEVELOPMENT

The following 17 generalized linear models with negative binomial distribution were developed using the Stata software. The results of all models were compared according to the assessment of model performance as detailed in section 2.5.4. In the first step, a model was developed only with a constant term and an appropriate offset. A stepwise incremental approach was followed in successive models by adding different variables. The Durbin-Watson test was used for the presence of serial correlation in the selected model. After this, a generalized estimation equation with AR1 error terms was estimated for the preferred model form, augmented by a lagged observation to allow for serial correlation. The coefficients and *t* values of the GLM and GEE-AR1 were then compared. The lattice of model development is shown in Figure 4.3. For each model that is fitted all the statistics are shown in Table 4.7.

4.5.1 Variables used

The following variables were used in development of the models:

1. Road class (five classes of road)

   Motorway │Rural A │Urban A │ Rural minor │Urban minor
2. Vehicle class (five classes of vehicle)

   Pedal cycle │ Motorcycle │Car │ Bus │Goods vehicle
3. Time (measured in days, with values from 1 to 1826, 1 January 2001 to 31 December 2005).
4. Logarithm (road length)
5. Day of week (with 7 levels)
6. Weekday 4 (4 levels: Weekday 1, Weekday 2, Saturday, Sunday)
7. Season (4 levels: Spring, Summer, Autumn, Winter)
8. Month of the year (12 levels)
9. Interaction of Weekday 4 and Season (16 levels)
10. Public holidays
11. Christmas holidays
12. New-Year holidays
13. Interaction of road class and vehicle class (With 24 levels)
14. Distance travelled per unit of road length
15. MC-Rural-Sunday (representing leisure motorcycling)

Here the categorical variable weekday 4 has 4 levels: weekday 1 which represents (Monday or Friday), weekday 2 (Tuesday, Wednesday or Thursday), Saturday, and Sunday. The details of this are given in section 4.6.1.1.

4.5.2 Basic structure of the model

In this chapter all models that were developed for Dataset 4 are shown in Figure 4.3. A measure of the total distance travelled on each day by road and vehicle class was used as an offset to represent the exposure to risk. This measure of distance was profiled by day of week and month to adjust the variations in distance travelled. As a result of this, the risk of road accident involvement per unit of this measure of distance can be estimated directly from the linear predictor.

In this study the number of vehicles involved in road accidents from each road and vehicle class has a panel structure with repeated observations: each road class and vehicle class combination (e.g. cars on motorway) corresponds to a member of the panel giving 24 combinations as shown in Table 4.11, each is measured repeatedly over the 1826 days of the study period. Dataset 4 has 43,824 observations and each observation represents the number of vehicles involved in road accidents occurring on each day for a member of panel.

The following forms of distance travelled were considered and tested for use as the basis of an offset in Dataset 4 models to represent the exposure to risk:

- Annual distance travelled
- Adjusted distance travelled on each day

The annual distance is available for each combination of vehicle class and road type. The distance travelled for each day was adjusted according to the day of week, month, by using the factors shown in Table 4.5. These adjustment factors were based on the road and vehicle classifications that are discussed in section 4.4. According to this, the distance travelled on each day will vary equally between all vehicle types (Pedal cycle, Motorcycle, Car, Bus and Goods vehicle) and equally on major and minor roads. The models that used adjusted distance travelled for each day as offset did not produce better goodness of fit in comparison to those that used annual distance travelled. The results in Table 4.6 show that *BIC* of model A1, in which the annual distance travelled is used as offset, is about 1,030 better than that of

model A2, in which adjusted distance travelled for each day is used. The use of annual distance travelled as offset for a model of numbers of vehicles involved in road accidents for each day overlooks the influence of day to day variation in distance travelled. Investigation of the effect of adjusting the distance travelled on each day to account for variation among days of the week and month of year led to reduced model performance. Notwithstanding this, because it is important to incorporate these variations (in the offset) so that it corresponds as closely as possible to the linear predictor for each unit of observation, the adjusted distance travelled for each day was adopted for use as offset. This also facilitates the interpretation of the coefficients as measures of risk per unit of distance of travelled.

Because of unobserved variables that affect the occurrence of road accident, we expect that there will be positive correlation among the numbers of vehicles of each of the classes that are involved in accidents on each day. This means that the single model that combines data from all combinations of road and vehicle class will have somewhat overstated likelihood and accuracy. Hence any coefficients that have marginal statistical significance are interpreted here with caution.

The following model structure was used for Dataset 4.

$$u_{ij} = \exp\left(O_{ij} + \mathbf{x}'_{ij}\,\boldsymbol{\beta}\right) \qquad 4\text{-}1$$

where *i* represents observation (corresponding to time) and *j* represents the member of the panel (combinations of road class and vehicle type),

u_{ij} is the estimated number of vehicles involved in road accidents occurring on each day *i* in road and vehicle combination *j*. and

O_{ij} is the offset: $O_{ij} = \ln\left(d_{ij}\right)$

Then

$$u_{ij} = d_{ij}\ \exp\left(\mathbf{x}'_{ij}\,\boldsymbol{\beta}\right) \qquad 4\text{-}2$$

where d_{ij} is the adjusted distance travelled on each day for observation i and road class and vehicle class combination j taking into account variation in distance travelled by day of week and month.

Table 4.6: Comparison of *BIC* with various measures of distance travelled used as offset

Model	**Annual distance travelled**		Model	**Adjusted daily distance travelled**	
	Log-likelihood	*BIC*		Log-likelihood	*BIC*
A1	-138,734	278,023	A2	-139,249	279,053
Variables used in Model A1 and A2					
Road class+ Vehicle type+ Time+ Weekday 4+ Season+ Month+ Weekday 4.Season + Public holidays+ Christmas holidays+ New-year holidays+ Road class.Vehicle type + MC-Rural-Sunday					

4.6 MODEL SELECTION PROCESS, GOODNESS OF FIT AND MODEL CHECKS

The model assessment procedure described in section 2.5.4 was applied to distinguish among many available models. The results of all the developed models shown in Figure 4.3 were compared. The details of all these models and various checks that were used to identify the appropriate model are given in sections 4.6.1.1 to 4.6.1.5.

4.6.1 Model Selection Procedure

The procedure outlined in section 2.5.4 was used to identify the preferred model out of the many that were developed to estimate the number of vehicles involved in road accidents occurring on each day. All of the models presented here were developed using GLM with negative binomial regression. The preferred model was then taken forward as the basis of investigation using the GEE formulation with autoregressive errors. The following section shows the results of this model selection procedure:

1. In section 4.6.1.1 the *BIC* values of the models are presented to compare their performance.
2. In section 4.6.1.2 the results of the analysis of temporal effects remaining in the models are presented.

3. In section 4.6.1.3 variance inflation factors are presented to check for the presence of multicollinearity in the data.
4. In section 4.6.1.4 split sample tests were carried out to validate the performance of the preferred model by comparing the coefficients, deviance and log-likelihood values.
5. In section 4.6.1.5 the presence of serial correlation in the preferred model was tested by using Durbin-Watson test.

4.6.1.1 Negative binomial regression model (Dataset 4)

A total of 17 models were developed as shown in Figure 4.3 using an incremental approach. An appropriate offset variable was used throughout this procedure, with adjustments introduced alongside corresponding explanatory variables.

In the first step a generalized linear model with negative binomial distribution was developed with only a constant term and using the logarithm value of distance travelled per day as offset. In models 2 and 3 road class and vehicle class variables were used individually. The reason for adding these terms into the model was to match the number of vehicles involved in road accidents on different road types and vehicle classes. In model 4 road class and vehicle type were used together. A continuous time variable was added in model 5. In model 6 the logarithm of road length in each class was introduced to investigate its effect on the model performance.

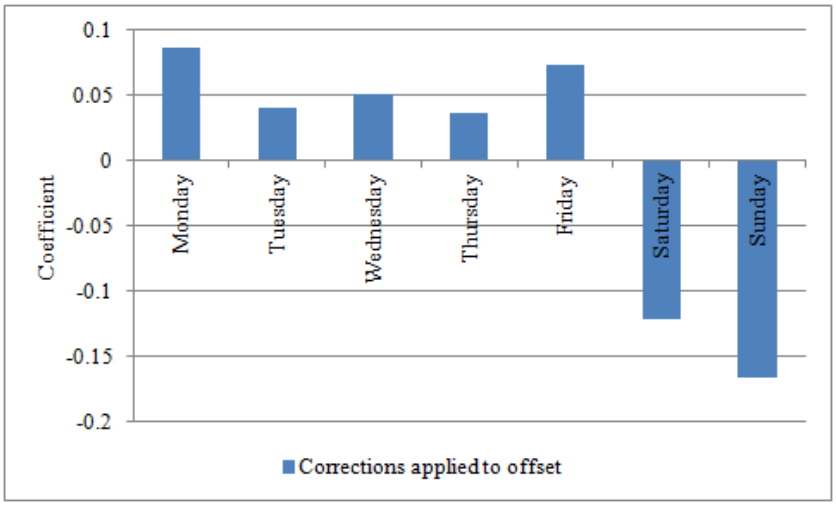
During the model development stages when weekday 4 was introduced into the linear predictor in model 7, the vehicle distance travelled in offset was profiled according to day of week by applying the corresponding correction factors obtained from Department for Transport to adjust the variation in vehicle distance travelled. Similarly when seasons in model 8 were included individually into linear predictor, the vehicle distance travelled (offset) was profiled accordingly. From model 9 onwards, the offset was profiled by day of week and month (correction factors of day of week and month were used together) when weekday 4 and seasons were used together into the linear predictor.

After this, weekday 4 variable representing each of weekday 1 (Monday, Friday), weekday 2 (Tuesday, Wednesday, Thursday), Saturday, and Sunday was used. Weekday 4 with 4 levels is simplified version of day of week with 7 levels. This variable was introduced instead of

day of week because it was observed from the graph shown in Figure 4.2 that when day of week variable was used along with the offset profiled only by day of week within model 7, the estimated coefficients which represent the risk per unit of travel were similar for Tuesday, Wednesday and Thursday. Monday and Friday also had almost same estimated coefficients as each other whereas those for each of Saturday and Sunday were substantially different. Due to this, weekday 4 variable was introduced instead of day of week in model 7.

Different explanatory variables including season, month, interaction of weekday 4 and season, Public holidays, Christmas holidays, New-Year holidays, the interaction variable of road class and vehicle class, and distance travelled per road length were used in models (8-16). Analysis of model 15 showed that observations belonging to motorcycle, Sunday and rural roads had particularly high deviance residuals. Motorcycling on rural roads on Sunday was considered as leisure activity. For this reason a special variable (MC-Rural-Sunday) was introduced in model 17 to separate the leisure motorcycling from other kinds of road use.

Figure 4.2: Comparison of the coefficients of day of week with different offset (Dataset 4)



For the first model, the *BIC* was found to be 360,422. It was found that road class (model 2) performed better than vehicle class (model 3) with *BIC* better by 12,926. After this, these two variables were used together in model 4 which improved the *BIC* value substantially, resulting in an improvement in *BIC* of 57,751 from model 1. Introduction of the Time variable resulted in an improvement of 861 in the *BIC* value of model 4 for one degree of freedom: model 5 had *BIC* of 301,810. The logarithm values of road lengths were introduced

in model 6. However, it was found that this variable did not improve the *BIC* in comparison to model 5, so this variable was not considered further.

In model 7, weekday 4 with 4 levels was introduced in place of the full 7 level day of week as explanatory variable and in model 8 season variable was introduced. It was observed that model 7 had better *BIC* values than model 8 suggesting that weekday 4 had performed better than season when used individually. In model 9, both of these variables were used together. The joint use of weekday 4 and season in model 9 had improved the *BIC* by 1,096 and 3,522 in comparison to model 7 and 8 respectively. Due to this model 9 with explanatory variables of road class, vehicle class, road class and vehicle class interaction, time, weekday 4 and month was taken forward.

Month variable was included in model 10 which improved the *BIC* by value of 72. Interaction of weekday 4 and season, Public holidays, Christmas holidays and New-Year holidays were also used in models 11 to 14 which also improved the performance, giving better *BIC* values. In model 15 interaction variables of road class and vehicle class were added, resulting in an improvement of 15,241 in *BIC* with an additional 15 degrees of freedom in comparison to model 14. Model 15 had a better *BIC* than all previous models with a value of 280,817.

In model 16 a new variable of distance travelled per unit of road length was introduced which reflected the usage of road class by vehicle class. This improved the *BIC* by 744 with one extra degree of freedom. This model was not considered further for the reasons that are explained in section 4.6.1.3. In model 17, a variable indicating leisure motorcycling was introduced. It was observed that the use of this variable was justified as *BIC* of the model improved by a value of 1,764 in comparison to model 15. Overall model 17 had the best results of all with an improvement of 81,369 (22 percent) in the value of *BIC* in comparison to model 1. The results of all 17 models are shown in Table 4.7.

Figure 4.3: Lattice of model development: Dataset 4

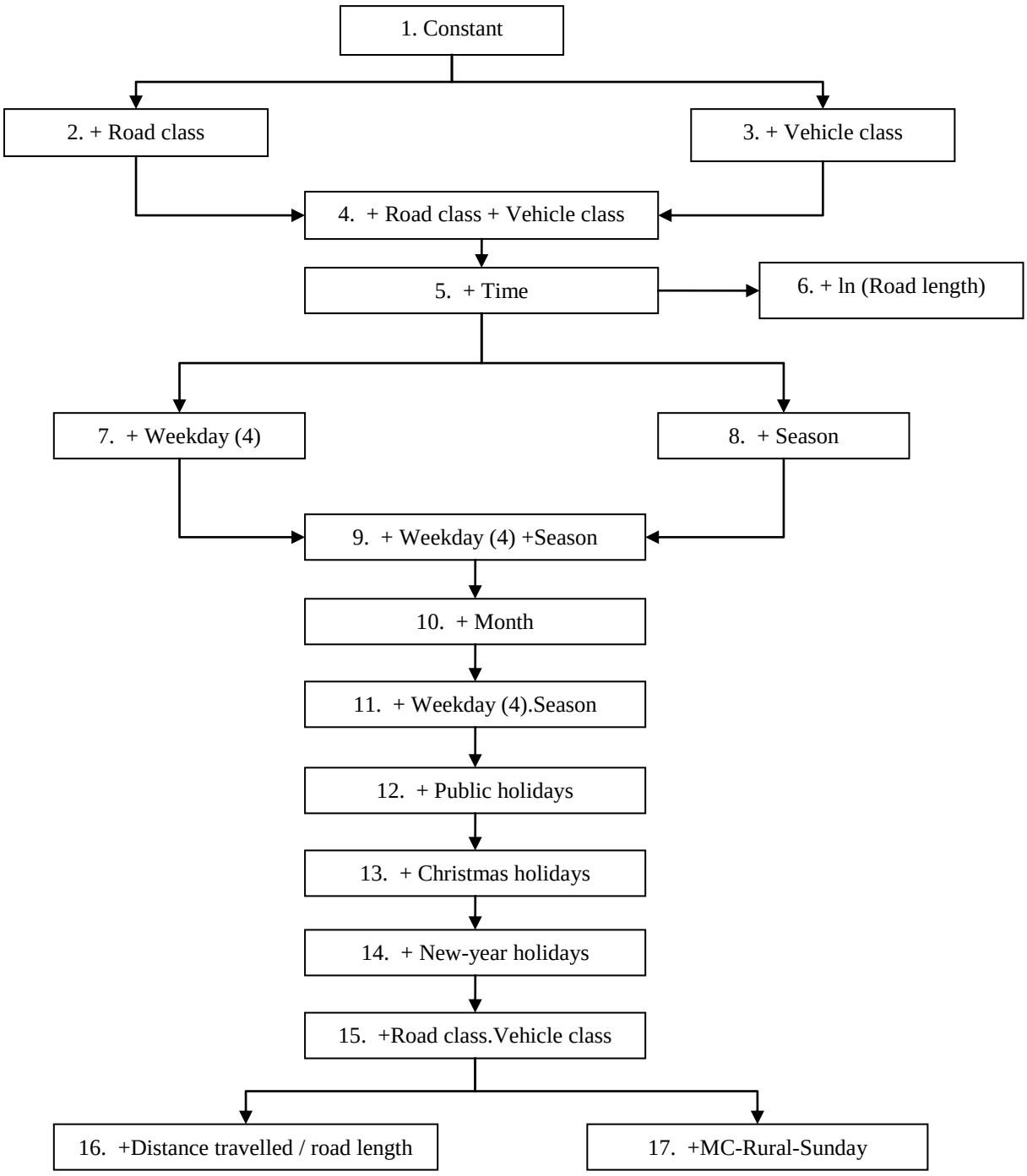


Table 4.7: Results of all models for each road and vehicle combination (Dataset 4)

Model	D.F	Scale	Likelihood	*BIC*
1	1	1.271	-180,205	360,422
2	5	0.602	-165,269	330,590
3	5	0.840	-171,731	343,516
4	9	0.247	-151,287	302,671
5	10	0.240	-150,852	301,810
6	11	0.240	-150,851	301,820
7	13	0.207	-149,147	298,432
8	13	0.233	-150,360	300,858
9	16	0.200	-148,583	297,336
10	24	0.199	-148,504	297,264
11	33	0.198	-148,359	297,070
12	34	0.192	-147,998	296,360
13	35	0.191	-147,910	296,194
14	36	0.190	-147,837	296,058
15	51	0.110	-140,136	280,817
16	52	0.105	-139,759	280,073
17	52	0.104	-139,249	279,053

BIC represents the Bayesian information criterion

4.6.1.2 Analysing the temporal effects

The developed models as shown in Figure 4.3 were analysed further to investigate any remaining substantial systematic temporal effect that was not represented in the model. For this purpose time and square of time variables were added to each of the models. The resulting improvement in *BIC*, coefficients and *t* values of time and square of time, and their variance inflation factors (VIF) were examined.

Here models 1-4 does not include time variable due to which time and square of time variables were added to those models. From model 5 onwards in which time variable was already present only the square of time was added to investigate the presence of substantial quadratic temporal effect.

Results presented in Appendix Table A4.1 showed that an improvement of over 290 in the value of *BIC* for models 2-4 (more than 800 in each of model 2 and 4) when time and square of time variables were added to the models. Comparatively small improvements in *BIC* from model 5 onwards were observed as time variable was already included into the models and had therefore represented most of the temporal trend. The *t* values of time and square of time were found to be significant in most cases, but the estimated value of VIF for each of time and square of time was in range of 16 which shows that these variables are correlated and their true effects can not be identified from the estimated parameters.

The most detailed model 17 showed that there is only improvement of 20 in the *BIC* when square of time is included. The *t* value of time and square of time was -5.76 and -5.53 respectively. However, high value of VIF shows that these variables are correlated with others. The small improvement in *BIC* of the models (5-17) in comparison to earlier models shows that any quadratic temporal trend in the data has been adequately represented by other variables in the model and only a small improvement in model performance can be achieved by allowing for further variation over time according to a quadratic term.

4.6.1.3 Checking for the presence of multicollinearity

Variance inflation factors as discussed in Chapter 2, section 2.6.2.3, were estimated for each of the models (6-17) to check for the presence of collinearity of the variables. The models with high VIF are less preferable.

In Table 4.8 mean values for road class and vehicle class are shown as representative of individual variables. Model 6 in which logarithm of road length was used, the VIF of road class is particularly high as a consequence of the structural collinearity between the road length and the road class variables. In models 7 and 8 where weekday 4 and season were used individually and in model 9 when these variables are used together had produced acceptable values of VIFs. From model 10, it was observed that season had collinearity with month. The VIFs of season arose due to the structural association with month so it was not a cause of concern. From models 11 to 14 it was found that VIF for the interaction variables of weekday and season, Public holidays, Christmas holidays and New-Year holidays were all within the acceptable range with values less than 3 in each model. The interaction variables of road class

and vehicle class in model 15 also produced high VIFs due to the structural relationship among the variables and so was ignored.

It is only when the variable of distance travelled per road length was used in model 16: the road class, interaction of road class and vehicle class, and distance travelled per road length had high VIF which showed collinearity between these variables. The VIF of the distance travelled per road length was 60.46. As of result of this, the true effects of these variables cannot be determined from the estimated coefficient, due to this model 16 was not preferred. Model 17 which had the better *BIC* than all other models and the new introduced variable of MC-Rural-Sunday (representing leisure motorcycling) was not correlated with any other variables hence this was preferred in comparison to other models and taken forward for further investigation. Table 4.8 shows the variance inflation factors of models 6-17 for Dataset 4.

Table 4.8: Variance Inflation Factors for Dataset 4

Model	Mean R.C	Mean V.C	Time	R.L	Mean WD_4	Mean Seas-ons	Mean Month	Mean WD_4. Season	Holida-ys	Christ-mas	New year	Mean R.C.V.C	D/L	M_R_S
6	51,299	1.67	1.04	109,315	-	-	-	-	-	-	-	-	-	-
7	1.67	1.67	1.00	-	1.25	-	-	-	-	-	-	-	-	-
8	1.67	1.67	1.02	-	-	1.57	-	-	-	-	-	-	-	-
9	1.67	1.67	1.02	-	1.25	1.57	-	-	-	-	-	-	-	-
10	1.67	1.67	1.04	-	1.25	13.95	5.10	-	-	-	-	-	-	-
11	1.67	1.67	1.04	-	1.33	14.33	5.10	2.12	-	-	-	-	-	-
12	1.67	1.67	1.04	-	1.33	6.05	4.69	2.11	1.06	-	-	-	-	-
13	1.67	1.67	1.04	-	1.34	6.40	3.67	2.11	1.27	1.27	-	-	-	-
14	1.67	1.67	1.04	-	1.34	6.38	3.67	2.19	1.48	1.3	1.21	-	-	-
15	7.79	7.79	1.04	-	1.34	6.14	4.74	2.12	1.48	1.3	1.21	17.84	-	-
16	97.53	9.68	1.04	-	1.37	6.09	4.69	2.11	1.48	1.3	1.21	72.28	60.46	-
17	7.79	7.79	1.04	-	1.39	6.14	4.74	2.12	1.48	1.3	1.21	17.91	-	1.26

Empty cells shows that these variables were not included in the corresponding models.

R.C=road class, V.C=vehicle class, R.L=road length, , WD 4= weekday 4, N-Y= New-Year holidays, D/L=distance travelled per road length, M_R_S=Motorcycle_Rural_Sunday

4.6.1.4 Split sample tests

After analysing the *BIC*, temporal effects and *VIF* values according to the criteria discussed in section 2.5.4, model 17 was taken forward for further investigation. Split sample tests were carried out on this model by randomly partitioning the whole of dataset 4 into two. Each part had 21,912 observations. The following datasets were used to cross-check and validate the results of model 17.

Full dataset = Data A
Dataset first portion = Data B
Dataset second portion = Data C

The Stata software was used to estimate the model parameters separately for model 17 using each of the Datasets B and C in turn. These were then compared with the coefficients of model 17 with the full data (Dataset A). After this, coefficients of model with Data B and C were interchanged to calculate the values of log-likelihood and deviance. This produced a small change in the original log-likelihood and deviance values. The coefficients of Dataset C when used with Dataset B produced likelihood of -69,575 which had a difference of only 31 from the value optimised for that dataset. Because the model parameters are not optimised in this case, there are 52 more degrees of freedom in the residuals; this gives rise to a likelihood ratio test statistic of 62 on 52 degrees of freedom, which is less than the critical value of 69.82 at 0.05 significance level. Therefore the null hypothesis cannot be rejected that the parameters fitted to Dataset C are as appropriate for Dataset B as those fitted to that dataset. In the same way, when coefficients of Dataset B were used with Dataset C that produced a difference of 27. As a result of this, the null hypothesis cannot be rejected that parameters fitted to Dataset B are as appropriate for Dataset C.

It was observed that results of the partitioned Datasets B and C do not differ widely. The most important finding is that when the coefficients of the partitioned data were exchanged it did not produce a large change in the results which indicates the consistency of the model. Together, these results presented in Table 4.9 show that the parameters of model 17 are consistent and produce approximately corresponding likelihood results.

Table 4.9: Split sample validation results for Dataset 4

Split sample validation				
		Model coefficients *(k=52)*		
Data		A	B	C
A		$\mathbf{x}'_A \boldsymbol{\beta}_A$		
	n	43,824		
	Likelihood	-139,249		
	Deviance	50,070		
B			$\mathbf{x}'_B \boldsymbol{\beta}_B$	$\mathbf{x}'_B \boldsymbol{\beta}_C$
	n		21,912	21,912
	Likelihood		-69,544	-69,575
	Deviance		25,072	25,247
C			$\mathbf{x}'_C \boldsymbol{\beta}_B$	$\mathbf{x}'_C \boldsymbol{\beta}_C$
	n		21,912	21,912
	Likelihood		-69,708	-69,681
	Deviance		25,141	24,995
Total	Likelihood	-139,249	-139,252	-139,256
	Deviance	50,070	25,213	50,242

In the second step of the validation process the coefficients of Datasets A, B and C are compared. The *T* test was used to compare the coefficients of Dataset B and C: T_{BC} values were estimated by using the formula 2-32. It is found that from the 52 variables used in model 17, only 1 changed significantly as its estimated *T* test value was greater than 1.96. All other variable except Bus did not change significantly. It is observed that coefficients of all variables and *t* values of the explanatory variables are consistent and carried the same sign in all three models. The comparison of coefficients and *t* values are shown in Table 4.10 and Figure 4.4. The following points were noted:

- The coefficient of road class and vehicle class had almost same coefficient and significant *t* values in all three models except Bus which was found to be non-significant in model B. Model A had more significant *t* values than model B and C.

- The coefficient of time was found to be negative and have a similar value of -0.066/ year in all three models. This corresponds to an annual reduction of about 6 percent in the number of vehicles involved in road accidents that caused personal injury.

- Each coefficient of weekday 4, season, month and interaction of weekday 4 and season was significant in each of the three models.

- The coefficient of public holidays, New-Year holidays, Christmas holidays were found to be significant in all three models.

- All the interaction variables of road class and vehicle class fitted were found to be significant in all three models except motorcycle on rural minor roads. Motorcycle on motorway was found to be non significant in model B only.

In summary, the split sample tests results showed good agreement between the parameter values estimated for model 17 based on two distinct subsets of the data. This stability supports use of the model, and the available parameter estimates from the model.

In this case deviation coding which is combination of (1, 0 and -1) is used to get the coefficients for factors that have zero mean for their effects. Due to this, coding structure the coefficient of Urban A will be equal to the minus sum of all other road classes. Same is for the coefficient of Car, Saturday, Spring, November and other variables. After this, the results were verified by comparing the likelihood values and estimated number of vehicles involved in road accidents by using simple coding (1 and 0) to further check that deviation coding has produced comparable results. These all coefficients estimated by using deviation coding are shown in Figure 4.4 and Table 4.10.

Table 4.10: Comparison of coefficient and *t* values of GLM-Model 17-NB for coefficient validation

	Comparison of the coefficients and t values of the Models						
	Model A		Model B		Model C		T test
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_c	T_{BC}
Motorway	-1.092	-146.35	-1.093	-103.94	-1.090	-103.00	-0.15
Rural A	-0.288	-40.96	-0.279	-28.15	-0.296	-29.79	1.21
Rural Minor	-0.160	-22.17	-0.157	-15.39	-0.162	-15.95	0.33
Urban Minor	0.725	104.76	0.724	73.75	0.725	74.37	-0.03
Pedal cycle	0.890	105.29	0.880	74.37	0.901	74.54	-1.25
Motorcycle	0.774	97.20	0.764	65.60	0.782	71.64	-1.13
Bus	-0.042	-4.89	*-0.020*	*-1.64*	-0.065	-5.28	2.60
Goods vehicle	-0.485	-59.84	-0.479	-41.69	-0.490	-42.87	0.68
Time	-0.00018	-43.32	-0.00018	-30.13	-0.00018	-31.08	0.48
Weekday 1	0.156	41.18	0.156	29.13	0.155	29.07	0.20
Weekday 2	0.107	31.29	0.110	22.68	0.104	21.64	0.84
Sunday	-0.197	-38.87	-0.202	-27.87	-0.192	-27.11	-0.98
Summer	0.104	11.89	0.098	7.81	0.111	8.97	-0.73
Autumn	-0.099	-5.88	-0.092	-3.84	-0.105	-4.44	0.38
Winter	0.035	4.21	0.040	3.39	0.030	2.59	0.62
January	-0.082	-7.80	-0.075	-5.04	-0.091	-6.07	0.74
February	-0.081	-7.64	-0.095	-6.28	-0.067	-4.48	-1.32
March	-0.065	-6.27	-0.065	-4.42	-0.065	-4.45	0.003
May	0.052	5.07	0.058	3.95	0.048	3.29	0.48
June	-0.050	-4.84	-0.038	-2.61	-0.060	-4.14	1.05
July	-0.066	-6.45	-0.058	-4.04	-0.073	-5.09	0.73
August	-0.073	-7.12	-0.070	-4.81	-0.075	-5.22	0.25
October	0.163	8.74	0.155	5.84	0.170	6.49	-0.40
WD1-Summer	-0.046	-8.06	-0.053	-6.48	-0.040	-4.95	-1.07
WD2-Summer	-0.043	-8.19	-0.039	-5.26	-0.046	-6.34	0.73
Sun-Summer	0.072	9.65	0.081	7.65	0.063	6.01	1.21
WD1-Autumn	0.033	4.55	0.034	3.28	0.031	3.11	0.15
WD2-Autumn	0.028	4.33	0.026	2.77	0.031	3.38	-0.39
Sun-Autumn	-0.039	-4.16	-0.042	-3.09	-0.036	-2.80	-0.32
WD1-Winter	0.033	5.18	0.033	3.62	0.033	3.69	-0.06
WD2-Winter	0.039	6.70	0.037	4.49	0.040	4.86	-0.24
Sun-Winter	-0.061	-7.30	-0.051	-4.28	-0.070	-5.96	1.12
Holidays	-0.153	-19.59	-0.153	-13.92	-0.154	-13.81	0.06
New-year	-0.329	-17.20	-0.324	-11.94	-0.333	-12.39	0.23
Christmas	-0.316	-14.16	-0.349	-10.38	-0.290	-9.67	-1.31
MC.Mot	-0.092	-3.35	*-0.045*	*-1.17*	-0.134	-3.49	1.63
Bus.Mot	-1.068	-19.94	-1.029	-14.06	-1.114	-14.17	0.79
GV.Mot	0.235	15.07	0.225	10.18	0.244	11.09	-0.63
PC.RA	0.735	35.07	0.727	24.77	0.742	24.82	-0.35
MC.RA	0.242	15.62	0.228	10.29	0.258	11.90	-0.98
Bus.RA	-0.645	-26.07	-0.662	-19.15	-0.629	-17.73	-0.65
GV.RA	0.251	17.52	0.248	12.30	0.254	12.43	-0.22
PC.RM	-1.039	-47.99	-1.036	-34.06	-1.042	-33.80	0.15
MC.RM	*-0.002*	*-0.13*	*0.005*	*0.19*	*-0.007*	*-0.29*	0.34
Bus.RM	-0.759	-27.30	-0.772	-19.92	-0.745	-18.68	-0.49
GV.RM	0.828	52.30	0.802	36.13	0.854	37.78	-1.64
PC.UM	-0.472	-32.22	-0.458	-22.16	-0.485	-23.39	0.93
MC.UM	0.140	10.02	0.160	7.98	0.122	6.21	1.34
Bus.UM	0.074	4.91	0.047	2.22	0.100	4.73	-1.76
GV.RM	1.081	76.19	1.064	53.17	1.097	54.52	-1.14

MC-Rural-Sun	0.950	42.35	0.912	27.73	0.978	31.89	-1.47
Constant	-14.347	-490.64	-14.376	-334.42	-14.326	-357.29	-0.86

Italic shows that these variables are not significant at 5 percent level.

Figure 4.4: Comparison of coefficients of model 17 using GLM-Negative Binomial (Dataset 4)

Month coefficient in the graph represents the combined effect of month and season

4.6.1.5 Durbin-Watson test

Because the dataset consists of a time series cross-sectional data, it is possible that serial correlation exists in the data, which could affect model estimates. The Durbin-Watson test was therefore carried out to investigate whether autocorrelation is present in the residuals. The presence of autocorrelation was tested in the whole dataset and in each combination of road class and vehicle class, which were considered to form a panel with five years' time-series data from 1st January 2001 to 31st December 2005. The observations for pedal cycles on motorway were excluded from the panel, which left 24 members, each with 1,826 observations. The formula given in equation 2-30 was used to calculate the Durbin-Watson Statistic, which was calculated for the whole dataset and for each panel member. The lower d_l and upper d_u critical values of 1.57 and 1.78 were obtained from Table 2.2 by using the number of observations and number of variables in the regression equation. If the estimated value was less than 1.57 the null hypothesis for the absence of autocorrelation was rejected and if the estimated value lay between 1.78 and 2.32 the null hypothesis was accepted. All other conditions either to accept, reject or inconclusive results are shown in Table 2.2. Based on the results of this test the null hypothesis of the absence of autocorrelation among residuals for the whole of dataset was rejected as the overall estimated value of Durbin-Watson statistic was 0.21 which was substantially less than critical value of 1.57. After this, the presence of autocorrelation was tested for each member of the panel. The hypothesis of the absence of autocorrelation among the residuals for each member of the panel was also rejected as the estimated value of Durbin-Watson statistic was less than d_l in each case. The overall results are shown in Table 4.11 which suggests that autocorrelation exists in each of the panel members so that its presence in the residuals should be considered.

Table 4.11: Durbin Watson test results for Dataset 4

Panel member	Name	DW	Panel member	Name	DW
2	Motor cycle. Motorway	0.02	14	Bus. Urban A	0.25
3	Car. Motorway	0.48	15	Goods vehicle. Urban A	0.18
4	Bus. Motorway	0.13	16	Pedal cycle. Rural Minor	0.15
5	Goods vehicles. Motorway	0.14	17	Motorcycle. Rural Minor	0.17
6	Pedal cycle. Rural A	0.12	18	Car. Rural Minor	0.29
7	Motorcycle. Rural A	0.19	19	Bus. Rural Minor	0.09
8	Car. Rural A	0.33	20	Goods vehicle. Rural Minor	0.15
9	Bus. Rural A	0.08	21	Pedal cycle. Urban Minor	0.23
10	Goods vehicle. Rural A	0.41	22	Motorcycle. Urban Minor	0.29
11	Pedal cycle. Urban A	0.21	23	Car. Urban Minor	0.41
12	Motorcycle. Urban A	0.29	24	Bus. Urban Minor	0.19
13	Car. Urban A	0.54	25	Goods vehicle. Urban Minor	0.25

4.6.1.6 Preferred model

Model 17 was preferred on the basis of the model assessment criteria discussed in section 2.5.4. The results showed that model 17 had better *BIC* values than all other models and the estimated values of VIF are also in acceptable range.

Other models were not preferred as their *BIC* was not better than model 17 or they had high VIFs. Model 16 was not preferred as its *BIC* was less preferable than model 17 (by value of 1,020) and the variables of road class and distance travelled per road length had high VIF, so that the true effect of these variables can not be identified. Model 15 was also not preferred as it had less preferable *BIC* value than model 17 (by value of 1,764). The residual analysis of model 15 also showed that this model had particularly high residuals for motorcycle, rural roads and Sunday. As a result of this the variable representing the motorcycling was

introduced in model 17 which improved the *BIC* and residual analysis in comparison to model 15.

The results of analysis of further temporal effects in section 4.6.1.2 also showed that in model 17 no substantial systematic temporal effect remains which can be accommodated by further quadratic temporal terms in the model. Split sample tests also verified that model 17 and its parameter estimates are consistent and reliable. Based on joint consideration of these and other model assessment criteria as described in section 2.5.4, model 17 was preferred. However, it was found that serial correlation existed in the data, so that GEE with AR1 error structure for model 17 was adopted to accommodate this. In the following section the coefficients of model 17 with GEE-AR1 and GLM with negative binomial are compared.

4.6.1.6.1 Comparison of coefficients for Dataset 4 (GEE-AR1 and GLM)

The GEE-AR1 model was used to estimate the coefficient and *t* values for model 17 by considering the data as a combination of panel and time-series data. The panel consisted of all combinations of road class and vehicle class. The correlation structure of autoregressive order 1 (AR1) for residuals was considered. A comparison was carried out between the coefficients and *t* values obtained by GEE-AR1 and GLM with negative binomial regression as shown in Table 4.12. Because the coefficients of these two models are estimated using the same data, they are not mutually independent so it is not immediately possible to test the differences between them. Instead they were compared informally. It is observed that coefficients of all variables are consistent and carried the same sign in both models. After comparing the *t* values estimated by these models it was found that generally the *t* values of the GEE-AR1 model were smaller than the GLM in most cases which suggests that the significance levels of these variables in the GLM model were inflated. However, the *t* values of weekday 1, Sunday, interaction of weekday 1 and summer, and interaction of weekday 1 and Autumn were found to be slightly higher in GEE-AR1 as compared to the GLM model. The coefficient of MC.RM (motor cycle on rural minor roads) was found to be non-significant in each of the models.

The coefficients of the variables presented here are arranged to have zero sum by deviation coding in STATA. Due to this, coding structure the coefficient of Car will be equal to the minus sum of all other vehicle classes. Same is for the coefficient of Urban A, Saturday,

Spring, November and other variables. In general it is found that Urban A roads had the greatest coefficient which shows higher risk per unit of distance travelled on these roads whereas motorways had the lowest coefficient indicating the least risk per unit of distance travelled. Pedal cycle and motorcycle have greatest risk per unit of travel in comparison to other vehicle types whereas Cars have the least risk. Weekday1 (Monday, Friday) had the greatest risk per unit of travel in comparison to weekday 2 (Tuesday, Wednesday, Friday) and each of Saturday and Sunday. Sunday had the least risk of vehicle involvement in road accident per unit of distance travel. Among the months of year September and November (combined effect of month and season) had the highest risk per unit of distance travelled whereas March had the least risk. The coefficients of Public holidays, Christmas and New-year holiday had negative sign which shows that fewer vehicles are involved in road accidents on these days, though it is not possible to assess risk on these days as no corrections are available for distance travelled.

The interaction coefficients which showed the additional effect for particular road and vehicle combinations highlighted that car on motorway, pedal cycles on A roads, bus on urban A roads, car on rural minor, goods vehicles on minor roads have higher risk than is suggested by the main effects. Similarly the interaction coefficients of Saturday and Sunday in spring and summer, weekday 1 and weekday 2 in autumn and winter had greater effects in addition to their main effects. The coefficient of leisure motorcycling (MC-Rural-Sunday) was found to be significantly positive. Because no specific correction could be made in the offset to distance travelled for this case, this coefficient can be taken to indicate a greater frequency of road accident involvement. However, in the absence of a suitable correction, no statement can be made about difference in risk per unit distance travelled.

In general the *t* values of coefficients in the GEE-AR1 and GLM with negative binomial were not same. This change suggests that if the presence of serial correlation in data is neglected then it may lead to incorrect inferences and could result in placing undue emphasis on those variables which are actually less significant. The comparison of the coefficients and their *t* values estimated using GEE-AR1 and GLM is given in Table 4.12 and Figure 4.5.

Table 4.12: Comparison of coefficients and t values of GEE-AR1 and GLM Model 17-NB for coefficient validation (Dataset 4)

Variables	Comparison of models			
	Model 17-GEE_NB		Model 17-GLM-NB	
	Coefficient	t_{GEE}	Coefficient	t_{GLM}
Motorway	-1.092	-128.97	-1.092	-146.35
Rural A	-0.288	-35.94	-0.288	-40.96
Rural Minor	-0.160	-19.48	-0.160	-22.17
Urban Minor	0.725	91.96	0.725	104.76
Pedal cycle	0.891	92.13	0.890	105.29
Motorcycle	0.773	84.99	0.774	97.20
Bus	-0.042	-4.21	-0.042	-4.89
Goods vehicle	-0.484	-52.17	-0.485	-59.84
Time	0.000	-38.14	-0.00018	-43.32
Weekday 1	0.155	43.68	0.156	41.18
Weekday 2	0.105	29.02	0.107	31.29
Sunday	-0.197	-40.10	-0.197	-38.87
Summer	0.106	10.70	0.104	11.89
Autumn	-0.095	-4.99	-0.099	-5.88
Winter	0.030	3.20	0.035	4.21
January	-0.079	-6.58	-0.082	-7.80
February	-0.076	-6.23	-0.081	-7.64
March	-0.063	-5.39	-0.065	-6.27
May	0.053	4.54	0.052	5.07
June	-0.051	-4.41	-0.050	-4.84
July	-0.066	-5.74	-0.066	-6.45
August	-0.074	-6.42	-0.073	-7.12
October	0.160	7.58	0.163	8.74
WD1-Summer	-0.047	-8.68	-0.046	-8.06
WD2-Summer	-0.043	-7.79	-0.043	-8.19
Sun-Summer	0.074	10.35	0.072	9.65
WD1-Autumn	0.033	4.85	0.033	4.55
WD2-Autumn	0.028	4.02	0.028	4.33
Sun-Autumn	-0.039	-4.30	-0.039	-4.16
WD1-Winter	0.034	5.67	0.033	5.18
WD2-Winter	0.039	6.24	0.039	6.70
Sun-Winter	-0.066	-8.03	-0.061	-7.30
Holidays	-0.146	-18.56	-0.153	-19.59
New-year	-0.300	-14.80	-0.329	-17.20
Christmas	-0.276	-12.16	-0.316	-14.16
MC.Mot	-0.093	-3.00	-0.092	-3.35
Bus.Mot	-1.075	-17.57	-1.068	-19.94
GV.Mot	0.238	13.36	0.235	15.07
PC.RA	0.730	30.65	0.735	35.07
MC.RA	0.248	14.08	0.242	15.62
Bus.RA	-0.644	-22.85	-0.645	-26.07
GV.RA	0.252	15.39	0.251	17.52
PC.RM	-1.047	-42.42	-1.039	-47.99
MC.RM	*0.004*	*0.20*	*-0.002*	*-0.13*
Bus.RM	-0.755	-23.90	-0.759	-27.30
GV.RM	0.830	45.87	0.828	52.30

PC.UM	-0.473	-28.29	-0.472	-32.22
MC.UM	0.140	8.74	0.140	10.02
Bus.UM	0.075	4.38	0.074	4.91
GV.RM	1.080	66.71	1.081	76.19
MC-Rural-Sun	0.899	41.07	0.950	42.35
Constant	-14.270	-469.19	-14.347	-490.64

(It is note that deviation coding is used in this case so the coefficient of the missing category (Example: Car) will be equal minus the sum of other vehicles (Pedal cycle, Motorcycle, Bus and Goods vehicle)

Figure 4.5: Comparison of coefficients of model 17 using GEE-AR1 and GLM Negative Binomial (Dataset 4)

Month coefficient in the graph shows the combined effect of month and season.

4.6.1.6.2 Comparison of the number of vehicles involved in road accidents observed and estimated, Standardised deviance residuals

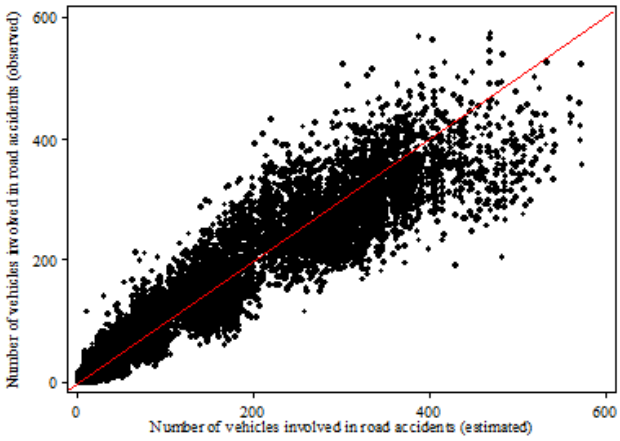
Model 17 was preferred over all others based on a joint consideration of *BIC* results, residuals analysis and estimated values of variance inflation factors. It was also found that serial correlation existed in the data due to which the GEE-AR1 was preferred over GLM.

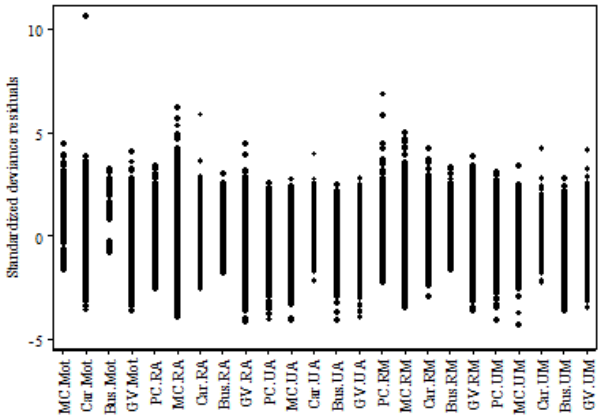
The graph in Figure 4.6 shows that model has generally represented the data well as the line of equality passes through the centre. The cumulative proportion graph shows no noticeable difference among the observed and estimated values.

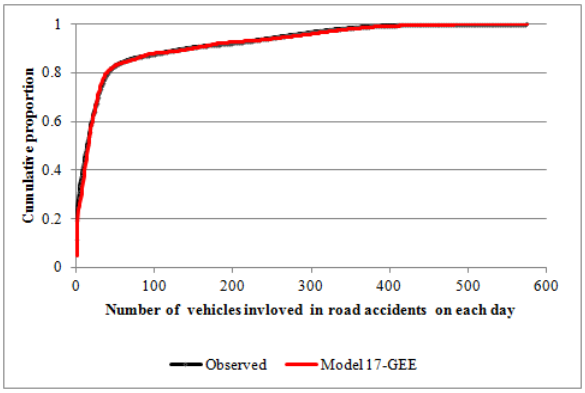
The standardized deviance residual graph also shows that the highest SDR observation (10.6) was for cars on motorways which occurred on 26th December 2004 which was Sunday. About 116 cars were found to be involved in road accidents on that day whereas the model estimated them as only 11. The estimated value for this observation was low because it was coded as Sunday, public holiday and Christmas holiday. After observing the data it was also found that the 27th and 28th December were also declared Public holidays (Monday and Tuesday) and, due to this long weekend travel, there might have been an increase in the amount of travel and subsequently an increase in observed road accidents. Upon further investigation it was observed that it snowed in many cities of Great Britain on 26th December (BBC, 2010).

Another high SDR observation was for pedal cyclist on rural minor roads on 17th June 2001 which was also Sunday. About 17 pedal cyclists were found to be involved in road accidents on rural minor roads but the model estimated only 1. It is generally observed that motorcycles on rural roads had a higher standardized deviance than all other groups. Out of the 100 observations with the highest positive SDR, 29 belonged to motorcycles on rural A roads while a further 17 belonged to motorcycles on rural minor roads. It is also found that most of these observations (42) related to Sundays. This suggests that the model is not able fully to capture this effect for motorcycles involved in a higher number of road accidents on rural roads especially on Sundays even after including the variable for the leisure motorcycling in model 17. Almost all the standardized deviance residual lies between the values of +5 and -5.

Figure 4.6: Number of vehicles involved in road accidents on each day (observed and estimated), Standardised deviance residual graphs (Dataset 4)







4.6.1.6.3 Final model checking

In this section, we investigate performance of model 17 fitted by GEE with negative binomial and AR1 error structure. To do this, some graphs are shown in Figure 4.7 to identify whether any problems exist in the model. The first graph shows the deviance residuals plotted against fitted values. It is observed that the plot of deviance residuals against fitted values appears to show some trend of falling variation with increase in estimated value. It was however found that about 67 percent of observations have estimated value (number of vehicles involved in road accidents by road type and vehicle class on each day) of less than 25 and that there is substantial variation in the density of observations over the range of fitted values. In particular, it was observed that the greatest residuals occur when the estimated number of vehicles involved in road accidents is under 10. The nature and strength of this variation in the deviance residuals was investigated by plotting in figure 4.8 averages of the absolute values of these residuals in bands of 50 of the estimated values. This graph reveals little trend in magnitude of deviance residuals though does suggest some positive curvature.

After this the Park and Glejser tests were used to investigate the presence of heteroscedasticity in the residuals. The test results shown in Appendix A4.2 confirmed the presence of heteroscedasticity. After this White's robust procedure was used to adjust the standard errors. We note that the hierarchical generalized linear model (HGLM) introduced and used in Chapter 5 allows to model variations in dispersion.

In table 4.13 the results of model 17 using GEE-AR1 are compared after adjusting the standard errors by using the White's procedure due to the presence of heteroscedasticity. The results show that *t* values of all the variables have decreased except for road class, vehicle class and their interaction. The coefficient of Winter turned to be non-significant after implementing the corrections to standard error whereas the coefficient of motorcycle on rural roads remained non significant in each case. This suggests that if the presence of heteroscedasticity is not accounted the coefficients will not be efficient but they will still be unbiased and consistent. Heteroscedasticity-corrected standard errors obtained by using the White's procedure are shown in Table 4.13.

In the second graph of figure 4.7, a normal quantile plot of standardized deviance residuals is shown. The quantile plot appears to follow a reference line except in the upper right portion.

This verifies the assumptions of normality of the residuals for most of the range of values. Some deviations are observed especially at the high end which suggests the data distribution has a long tail at that end.

Cook's distance plot shows the observations that had greater influence on the results. The highest Cook's distance was observed for 26th December 2004 for cars on motorways. This was a Sunday and the number of cars involved in road accidents was 116 against an estimated value of only 10. However, the value of Cook's distance was less than 0.1, showing that this observation did not have an undue effect on model estimates.

Figure 4.7: Diagnostic plots for model 17 (Dataset 4)

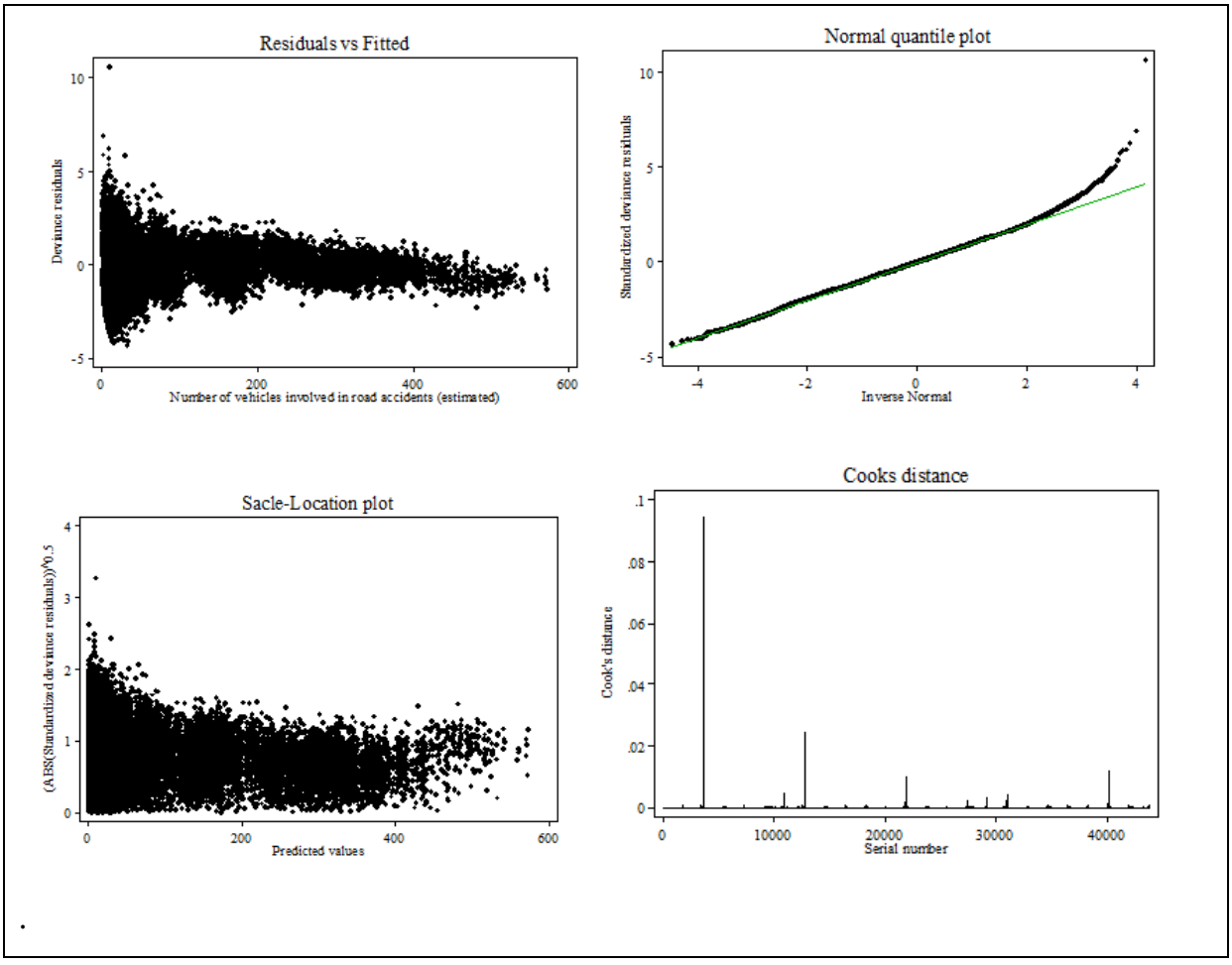
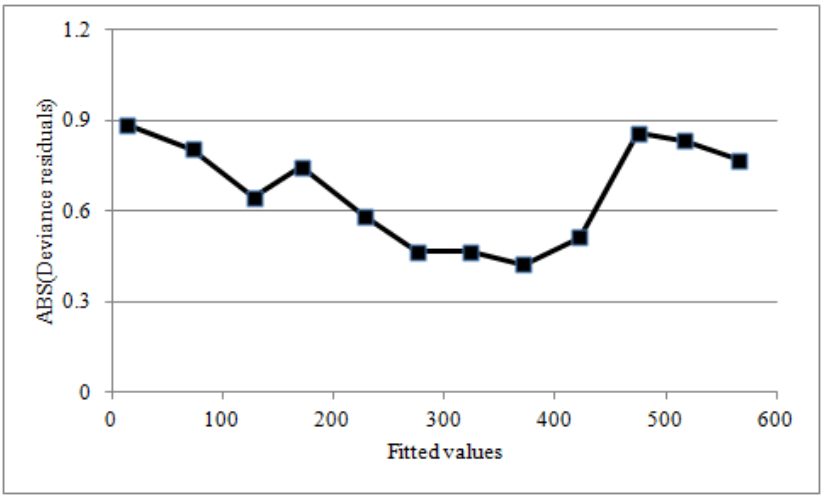


Table 4.13: Comparison of coefficient and *t* values of GEE-AR1 Model 17-NB after using correction for the presence of heteroscedasticity

Variables	Comparison of models			
	GEE-AR1 Model 17		GEE-AR1 Model 17-Robust	
	Coefficient	t_{GEE}	Coefficient	T_{GEE}
Motorway	-1.092	-128.97	-1.092	-780.04
Rural A	-0.288	-35.94	-0.288	-987.06
Rural Minor	-0.160	-19.48	-0.160	-401.18
Urban Minor	0.725	91.96	0.725	940.35
Pedal cycle	0.891	92.13	0.891	679.04
Motorcycle	0.773	84.99	0.773	6524.29
Bus	-0.042	-4.21	-0.042	-35.51
Goods vehicle	-0.484	-52.17	-0.484	-3394.11
Time	-0.00018	-38.14	-0.00018	-4.47
Weekday 1	0.155	43.68	0.155	3.94
Weekday 2	0.105	29.02	0.105	2.24
Sunday	-0.197	-40.10	-0.197	-2.98
Summer	0.106	10.70	0.106	3.17
Autumn	-0.095	-4.99	-0.095	-5.14
Winter	0.030	3.20	*0.030*	*0.67*
January	-0.079	-6.58	-0.079	-3.95
February	-0.076	-6.23	-0.076	-3.40
March	-0.063	-5.39	-0.063	-2.76
May	0.053	4.54	0.053	3.63
June	-0.051	-4.41	-0.051	-3.74
July	-0.066	-5.74	-0.066	-5.09
August	-0.074	-6.42	-0.074	-5.89
October	0.160	7.58	0.160	5.76
WD1-Summer	-0.047	-8.68	-0.047	-3.60
WD2-Summer	-0.043	-7.79	-0.043	-3.59
Sun-Summer	0.074	10.35	0.074	3.76
WD1-Autumn	0.033	4.85	0.033	3.90
WD2-Autumn	0.028	4.02	0.028	3.26
Sun-Autumn	-0.039	-4.30	-0.039	-2.65
WD1-Winter	0.034	5.67	0.034	2.78
WD2-Winter	0.039	6.24	0.039	3.02
Sun-Winter	-0.066	-8.03	-0.066	-3.41
Holidays	-0.146	-18.56	-0.146	-3.78
New-year	-0.300	-14.80	-0.300	-4.62
Christmas	-0.276	-12.16	-0.276	-4.82
MC.Mot	-0.093	-3.00	-0.093	-59.63
Bus.Mot	-1.075	-17.57	-1.075	-501.54
GV.Mot	0.238	13.36	0.238	96.31
PC.RA	0.730	30.65	0.730	195.00
MC.RA	0.248	14.08	0.248	26.85
Bus.RA	-0.644	-22.85	-0.644	-441.58
GV.RA	0.252	15.39	0.252	95.69
PC.RM	-1.047	-42.42	-1.047	-588.94
MC.RM	*0.004*	*0.20*	*0.004*	*0.41*
Bus.RM	-0.755	-23.90	-0.755	-636.96
GV.RM	0.830	45.87	0.830	591.26

PC.UM	-0.473	-28.29	-0.473	-118.13
MC.UM	0.140	8.74	0.140	47.16
Bus.UM	0.075	4.38	0.075	30.10
GV.RM	1.080	66.71	1.080	430.66
MC-Rural-Sunday	0.899	41.07	0.899	9.67
Constant	-14.270	-469.19	-14.270	-108.05

Figure 4.8: Average of the absolute value of deviance residual and estimated values in bands (Dataset 4)



4.7 Estimation of Risk per Vehicle kilometre of Travel

The risk of vehicle involvement in road accident per vehicle kilometre of travel is estimated by using the procedure shown below. The numbers of vehicles involved in road accidents estimated by model 17 and distance travelled adjusted by day of week and month corrections was used to estimate the risk per billion kilometres of travel for different road and vehicle combinations.

4.7.1 Estimating the number of vehicles involved in road accidents

In the first step, the average number of vehicles involved in road accidents on each day for the 24 combinations of road class and vehicle type were estimated from the observed and estimated (model 17) data. The results in Table 4.14 shows that the estimated values for the average number of vehicles involved in road accidents on typical day for each road and vehicle type closely matched with the observed numbers of vehicles.

It is found from the estimated values that on urban A roads an average of 260 cars were involved in road accidents per day whereas on motorways cars were involved in fewer road accidents than on all other roads with an average of 41 accidents per day. On urban minor roads an average of 340 cars were involved in road accidents. Motorcycles were involved in road accidents on each day equally on urban roads with average of 29 on urban minor and 26 on urban A roads. Comparatively, very few motorcycles were involved in road accidents on motorways. Pedal cycles were involved in fewer road accidents on rural roads in comparison to urban roads.

The highest incidence of pedal cycles in road accidents was observed on urban minor roads with an average of 29 road accidents whereas an average of fewer than three pedal cycles were involved in road accidents on rural roads. Buses were also involved on average in very few road accidents on motorways and rural roads. The results show that buses will be involved in one accident for every six days on motorways. After cars, motorcycles are hugely involved in road accidents on urban A roads.

It was found that on each of the urban A and urban minor roads an average of more than 20 goods vehicles per day were involved in road accidents compared with only 7 on rural minor roads, and 10 on motorways. The detailed results of the estimated number of vehicles involved in road accidents for all road and vehicle combinations is shown in Table 4.14 which shows that on average 436 vehicles will be involved in road accidents on urban minor roads, of which 77 percent will be cars. In the same way an average of 49 pedal cycles and 72 motorcycles were involved in road accidents on all roads with the majority of these occur on urban roads.

Table 4.14: Estimated risk per billion vehicle kilometres of travel and number of vehicles involved in road accidents per day estimated by model 17 GEE-AR1 (NB)

Vehicle class	Road classification					
	Motorway	A Roads		Minor		Overall risk by vehicle type
		Rural	Urban	Rural	Urban	
Pedal cycle						
Risk	-	4,621	10,760	864	3,771	4,180
Observed	-	2	16	2	28	48
Estimated	-	(2)	(16)	(2)	(29)	(49)
Motorcycle						
Risk	797	2,830	9,517	2,509	6,133	5,026
Observed	1	10	27	6	28	72
Estimated	(1)	(10)	(26)	(6)	(29)	(72)
Car						
Risk	211	471	1,414	532	1,303	793
Observed	40	136	252	70	326	824
Estimated	(41)	(141)	(260)	(72)	(340)	854
Bus						
Risk	133	454	4,246	458	2,547	2,070
Observed	0.18	1	14	1	13	29
Estimated	(0.18)	(1)	(14)	(1)	(13)	(29)
Goods vehicle						
Risk	316	714	2,717	1,436	4,527	1,063
Observed	10	20	22	7	27	84
Estimated	(10)	(20)	(22)	(7)	(25)	(84)
Overall risk by road class						
Risk	228	521	1,696	597	1,534	
Observed	52	169	331	85	422	
Estimated	(52)	(174)	(338)	(86)	(436)	

- Risk represents the risk of road accident per billion vehicle kilometres of travel.

4.7.2 Estimation of risk of an accident per billion vehicle kilometres of travel

After estimating the number of vehicles involved in road accidents for each road class by vehicle type on each day, the risk per unit of travel was estimated by dividing by the respective traffic flow. The detailed results of the estimated risk are shown in Table 4.14.

This shows that although cars were involved in huge numbers of road accidents, the risk of involvement per billion vehicle kilometres of travel was lowest for cars on all roads except for motorways and rural roads where buses are safer. These results suggest that pedal cycles were at higher risk on A roads whereas on minor roads the risk for motorcycles was higher than all other modes. The risk for pedal cycles on A roads is alarming especially on urban A roads with 10,760 pedal cycles involved in road accidents per billion vehicle kilometres of travel. Motorcycles, despite having higher involvement in road accidents on urban minor roads than urban A roads had a lower risk per unit of travel on urban minor roads.

Cars were found to have lower risk on motorways than other kinds of road. It was also found that although the number of buses involved in road accidents was almost same for urban A and urban minor roads the risk of a bus being involved in road accident on urban A roads was 66 percent higher than on urban minor roads. Goods vehicles were also at more risk on urban minor roads with 4,527 road accidents per billion vehicle kilometres of travel. On A roads, pedal cycles and motorcycles had a higher risk than any other mode on the same kind of road.

The risk of involvement in road accidents for different vehicle classes was compared with others, the details of which are given as follows:

4.7.2.1 Comparison of the risk per billion vehicle kilometres for pedal cycles with other vehicle classes

Table 4.15 shows the comparison of the risk between vehicle classes. It shows that:

- On rural and urban A roads pedal cycles had at least a seven times higher risk of involvement in a road accident than a car.
- Motorcycles had less risk on minor roads than on major roads. They had about three times higher risk than pedal cycles on rural minor roads.
- On rural A roads pedal cycles had at least ten times higher risk of a road accident than buses.

- Pedal cycles were at six times more risk than goods vehicles on rural A roads whereas on minor roads goods vehicles were at a higher risk than pedal cycles.

4.7.2.2 Comparison of the risk per billion vehicle kilometres of motorcycles with other vehicle classes

- Motorcycles had a low risk per unit of travel on A roads in comparison to pedal cycles whereas they had a higher risk than pedal cycles on minor roads.
- Motorcycles had at least six times higher risk than cars on A roads whereas on minor roads the risk was four times higher risk than for cars.
- On motorways and rural roads, motorcycles had around six times higher risk than buses. On urban roads the risk was about two times greater than for buses.
- On A roads motorcycles had about four times higher risk than goods vehicles.

4.7.2.3 Comparison of the risk per billion vehicle kilometres of cars with other vehicle classes

- Generally cars were safer on all roads than all other modes of transport except buses on motorways and rural roads. On motorways the risk of car being involved in accidents was 60 percent more than for a bus.
- On rural A roads cars had about the same risk of road accident than bus.

4.7.2.4 Comparison of the risk per billion vehicle kilometres of buses with other vehicle classes

- Buses had a lower risk than most other vehicles on all types of road.
- Buses had about 50 percent more risk than goods vehicles on urban A roads.
- On urban roads buses had about two times higher risk than cars.

Table 4.15: Comparison of risk per billion vehicle kilometres between vehicle types

Road class	PC	MC	Car	Bus	GV
Motorway		-	-	-	-
Rural A	*Reference*	1.63	9.81	10.18	6.47
Urban A		1.13	7.61	2.53	3.96
Rural Minor		0.34	1.62	1.89	0.60
Urban Minor		0.61	2.89	1.48	0.83
	PC	**MC**	**Car**	**Bus**	**GV**
Motorway	-		3.72	5.99	2.52
Rural A	0.61	*Reference*	6.01	6.23	3.96
Urban A	0.88		6.73	2.24	3.50
Rural Minor	2.90		4.72	5.48	1.75
Urban Minor	1.63		4.71	2.41	1.35
Road class	**PC**	**MC**	**Car**	**Bus**	**GV**
Motorway	-	0.26		1.59	0.67
Rural A	0.10	0.17	*Reference*	1.04	0.66
Urban A	0.13	0.15		0.33	0.52
Rural Minor	0.62	0.21		1.16	0.37
Urban Minor	0.35	0.21		0.51	0.29
	PC	**MC**	**Car**	**Bus**	**GV**
Motorway	-	0.17	0.63		0.42
Rural A	0.10	0.16	0.96	*Reference*	0.64
Urban A	0.39	0.45	3.00		1.56
Rural Minor	0.53	0.18	0.86		0.32
Urban Minor	0.68	0.42	1.95		0.56
	PC	**MC**	**Car**	**Bus**	**GV**
Motorway	-	0.40	1.51	2.38	
Rural A	0.15	0.25	1.52	1.57	*Reference*
Urban A	0.25	0.29	1.92	0.64	
Rural Minor	1.66	0.57	2.70	3.14	
Urban Minor	1.20	0.74	3.47	1.78	

4.7.2.5 Comparison of the risk per billion vehicle kilometres of goods vehicles with other vehicle classes

- Goods vehicles had a lower risk than pedal cycles on A roads but a greater risk on minor roads.

- Goods vehicles had less chance of being involved in accidents than motorcycles on all types of road.
- On each type of road, goods vehicles had a higher risk than cars especially on urban minor roads where they had a three times higher risk than cars.
- Goods vehicles had a lower risk than buses on urban A roads but had a higher risk than buses on all other roads especially on rural minor roads where they had a three times higher risk than buses.

4.8 CONCLUSION

The purpose of this chapter was to use the road accident dataset in a better way by combining the accidents and vehicle sections of the STATS 19 data. A further objective was to formulate a model from the national road accident dataset to estimate the number of vehicles involved in road accidents occurring on each day by type of road and by vehicle class, which can be used by planning and road safety organizations for improving road safety. These results will also support advice to travellers and can be used for education and increasing awareness about the groups that are at most risk per unit of travel.

It was found that in this case serial correlation exists in the data used in modelling, arising from its nature as a time-series. In order to draw inferences from such models for policy or road safety improvement purposes a suitable method should be applied which can account for the serial correlation, otherwise it may lead to incorrect inferences. In this case better performance was achieved by GEE-AR1 than the GLM for the estimated number of vehicles involved in road accidents. Difference, especially in levels of significance was found between GLM and the preferred GEE-AR1 model.

Several effects have been identified and discussed that would weaken a statistical model of numbers of vehicles involved in road accidents based on an independent Poisson error structure. These include over-dispersion, serial correlation, day to day variation in distance travelled and correlation between the numbers of vehicles in different classes involved on each day. Of these, over-dispersion was accommodated using the negative binomial error structure, serial correlation was addressed using the GEE model formulation with AR1 error structure and day to day variation in distance travelled was incorporated by using the corresponding correction factors to the offset. However, lack of allowance for the correlation

among members of the panel remains a limitation to the model that will lead to overestimation of the significance level of estimated parameters. Due to this, the coefficients that are marginally significant are treated with caution.

In this case the distance travelled each day was adjusted to account for variation by day of week and month of the year. This was preferred for use as offset in comparison to use of annual average distance travelled because the associated model coefficients can be interpreted directly in terms of risk per unit of travel. From the modelling results it is also observed that use of road class, vehicle type, and the interaction variable of road class with vehicle type greatly improved the performance of the model.

From the estimated results it is found that each of Monday and Friday has greater risk of vehicle involvement in road accident per unit of distance travelled than other days of the week. Weekends days in particular are associated with lower risk. November and September had greater risk whereas March had lowest risk among the month of year. Time variable showed that the risk of vehicle involvement in road accident per unit of travel is decreasing annually by about 6 percent. Fewer vehicles are involved in road accidents on Public holidays, Christmas and New-year holiday, though in the absence of appropriate adjustments to distance travelled on these days, nothing can be said about risk.

Urban roads had the greater risk of road accident than other roads. Motorways were found to have less risk per unit of distance travelled for all user classes. It is concluded that cars are involved in more road accidents than any other vehicle class. Despite their huge involvement in accidents the risk per billion vehicle kilometres for cars is low on all road classes in comparison to other vehicles classes except buses on motorways and rural roads. Motorcycles are at more risk than any other vehicle class on motorways and on minor roads, whereas pedal cycles are at more risk than any other vehicle class on A roads, whether urban or rural. It is also found that leisure motorcycling is associated with greater frequency of involvement in road accidents than other forms of motorcycle usage, though it was not possible to assess risk as no corrections are available for distance travelled. It is also concluded that cars, motorcycles, pedal cycles, and buses are at a higher risk of accident involvement on urban A roads in comparison to all other roads whereas goods vehicles are at most risk on urban minor roads.

5. MODELLING THE NUMBER OF CASUALTIES IN ROAD ACCIDENTS

5.1 INTRODUCTION

It is well known that age group and gender have a high relevance to road safety. In Great Britain young drivers aged between 17 and 24 years old are considered to be a high risk group in terms of road casualties. Although this group represents only 8 percent of driving licence holders nationally, they contribute to 20 percent of all driver casualties. On the other hand older motorists bring a wealth of experience, confidence, and tolerance to their driving which contributes to making them safer per licence holder on the road than other age groups. However with increasing age, ability to interpret the movements and intentions of other drivers and reaction time to different situations gradually changes. The physical body strength also changes and older age people are less likely to survive the injuries which a young person can survive (NCC Road safety, 2006).

The risk per unit of travel of being involved in road accident may vary with age and gender. According to the Department for Transport (2004a, 2004b) within adults, the risk of being involved in pedestrian accident varies with age and gender, with older adults at greatest risk of being seriously injured or killed per distance walked and men at all ages being at greater risk of serious injury than women. The UK Government set targets to reduce the number of casualties to a certain level by 2010 in comparison to base 1994-1998 average. The DfT (2011) revealed that all the targets have been achieved. The key results produced by the DfT (2011) are:

- 25,845 pedestrian casualties occurred in 2010 which was 44 percent lower than in 1994-1998 average.
- 17,185 pedal cyclist casualties occurred in 2010 which was 30 percent lower than compared to 1994-1998 average.
- 18,686 motorcycle user casualties occurred in 2010 which was 22 percent less than 1994-1998 average.
- 133,205 car user casualties occurred in 2010 which was 34 percent lower than 1994-1998 average.

- 208,648 road casualties occurred in 2010 which was 35 percent lower than 1994-1998 average.

The aim of this research is to explore further possibilities for the use of national accident data in conjunction to other available data. In previous chapters information from the accident and vehicle sections of STATS 19 data was used. In this section, combined information from the accident and casualty sections of STATS 19 data was used. As the information about the age group and gender only appears in the casualty section of the STATS 19 data, the accident and casualty sections of STATS 19 data were combined by extensively using MS Access and SPSS. This new combined dataset will be used to link the two separate sections of the STATS 19 data. The other datasets which were combined with the accident and casualty data include National Travel Survey data (NTS) obtained from DfT and population datasets produced by Office for National Statistics, United Kingdom.

This research has following objectives;

- investigate the relationships in the casualty data;
- investigate casualty data using the Hierarchical Generalized Linear Model (HGLM) to see what additional structure in the data is revealed;
- quantify any bias in estimates of coefficients estimated using simpler models such as GEE; and
- estimate the casualty rate of involvement in a road accident per person-years for different age and gender groups by vehicle class.

The HGLM is an extension of Generalized Linear Model (GLM) which allows for the fixed effects, as does the GLM, but in addition allows for random effects and a structured variance model for dispersion. The advantage of HGLM that it can account for variability within and between clusters using both random effects and dispersion modelling provided a substantial advantage over GLM and GEE. However, HGLM cannot accommodate time series data due to which the significance levels of some of the variables may change significantly.

This study will identify a suitable technique for modelling the number of casualties occurring on each day from the national accident dataset by highlighting the additional modelling

benefits of using HGLM. The number of casualties, disaggregated by day of week, month, year, age group, gender, and mode combination for Great Britain from 2001 to 2005 extracted from the STATS 19 national accident dataset were modelled and compared with the casualties which actually occurred. This comparison will enable researchers working in the field of road safety to understand the relationship between the number of casualties and other variables particularly age group, gender and mode. From the estimated number of casualties the rate of being a road casualty per head of population can also be estimated. Models were initially developed by using HGLM with a Poisson-gamma distribution and log link. For the selected model the Generalized Estimation Equation (GEE-AR1 error structure) with negative binomial was used and results are compared with HGLM. The estimated rate values per head of population for all age groups, gender, and mode combinations can be utilised to create awareness for any target group. The identification of the target group will help various planning agencies to have a clear picture of the number of casualties and rate per head of population by age group, gender, and mode which may enable the respective authorities to focus on a particular group and plan road safety schemes for targeted groups.

This chapter is organized as follows. Section 5.2 reviews the literature about the hierarchical generalized linear model and previous research about road accidents by age, gender and mode of travel. Section 5.3 briefly describes the data used for this study. Section 5.4 briefly analyses the data. Section 5.5 presents the process of model development and the basic structure of the model. Section 5.6 shows the model selection process, results of developed models, goodness of fit and model checks. Finally some concluding remarks are given in Section 5.7.

5.2 LITERATURE REVIEW

In the present study the Hierarchical Generalized Linear Model (HGLM) with Poisson-gamma distribution and log link, and the Generalized Estimation Equation (GEE) having AR1 error structure with negative binomial were used. The description of the GEE is given in Chapter 2 whist the HGLM is described below:

5.2.1: Fixed and random effects

There are various applications where it is believed that responses depend on some factors, but not all of which are known or measurable. Such unknown variables can be modelled as random effects. In case of repeated measurements for a subject, a random effect is an unobserved variable for each subject that is responsible for creating the dependence between repeated measures. Random effects may be regarded as a sample from a suitably defined population (Grafen and Hails, 2002; Lee et al, 2006). This differs from fixed effects, whose levels are of interest in their own right. Desired inference and repetition are the two properties which are most used to distinguish fixed from random effects. In the case

$$Y = \text{fixed effects} + \text{error} \qquad 5\text{-}1$$

the variance in Y is the sum of variance partitioned between that which is explained by the fixed effects and that which remains unexplained. On the right hand side of equation 5.1, only the error term has random variation which means it is the only term which will vary in repetitions of the study. The error term also determines the independence of each observation. The main assumption of the Generalized Linear Model (GLM) is that error terms are mutually independent. However in the presence of random effects the relevant equation is:

$$Y = \text{fixed effects} + \text{random effects} + \text{error} \qquad 5\text{-}2$$

In this equation the random effects term also has random variation. If the random effects term is unimportant, then estimated parameters of this factor will be close to zero and this term vanishes from equation 5-2. However, if the random effect term is important it will lead to the conclusion that individuals or subjects are different from each other. In this case, variation is divided into parts by separating the variation due to random effects and that due to the error term. According to Lee et al (2006) fixed effects describe systematic mean patterns such as trend, while random effects may describe either the correlation patterns between repeated measures within subjects or heterogeneities between subjects or both. In estimating a random effect, the observed deviations are characterised by their variance.

5.2.2 Hierarchical generalized linear model (HGLM)

The HGLM is the extension of the GLM and the Generalized Linear Mixed Model (GLMM). Pierce and Sands (1975) introduced the GLMM where the linear predictor of the GLM is allowed to have, in addition to usual fixed effects, one or more random components with assumed normal distributions. Lee and Nelder (1996) extended GLMM to HGLM, in which the distribution of random components is extended to conjugates of arbitrary distributions from the exponential family. The HGLM approach provides a unified modelling framework for estimating cluster-specific quantities of interest, covariate effects, and components of variance. These models make precise estimates of case-specific and cluster-specific parameters. They also produce reliable standard error estimates which are more realistic than the models in which random effects are not taken into consideration. One of the advantages of HGLM is the joint modelling of mean and dispersion. Dispersion parameters are allowed to have structures defined by their own set of covariates. It is useful to build a complex model by combining component GLM. The complete model is then decomposed into several components which provide additional insights into the model (Lee et al, 2006).

In general, the following are the three major benefits of using HGLM:

1. Heterogeneity between clusters, which is associated with unequal variances and arises from various sources, can be modelled by introducing a random effect into the mean model;
2. HGLM can be used to account for variability within and between subjects; and
3. Dispersion can also be modelled and significance of the variables in the dispersion model can be tested.

Lee, Nelder and Pawitan (2006, p173) define the HGLM as:

1. Conditional on random effects u, the responses y follow a GLM family, satisfying

$$E(y \mid u) = \mu \qquad \textit{(Lee et al., 2006, 173, ff)} \qquad 5\text{-}3$$

$$V\mathrm{ar}(y \mid u) = \phi V(\mu) \qquad 5\text{-}4$$

where $V(\mu)$ is the variance function, for which the kernel of the likelihood is given by

$$\sum\{y\theta - b(\theta)\}/\phi. \qquad 5\text{-}5$$

The parameter θ, which can vary according to u, is known as the *canonical* parameter. The linear estimator takes the form

$$\eta = g(\mu) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{v}\mathbf{Z}', \qquad \textit{(Lee et al., 2006, 174, ff)} \qquad 5\text{-}6$$

where $\mathbf{v} = \mathbf{v}(\mathbf{u})$ for some monotone function $\mathbf{v}(\,)$ represents the random effects with model matrix **Z,** and $\boldsymbol{\beta}$ are the fixed effects.

2. The random component u follows a distribution conjugate to a GLM family of distributions with parameters λ.

5.2.3: Basic structure of the HGLM

In the HGLM model formulation that is adopted here, the distribution of $y|u$ is Poisson with mean

$$\mu = E(y|u) = \exp(\mathbf{x}'\boldsymbol{\beta})u \qquad \textit{(Lee et al., 2006, 174, ff)} \qquad 5\text{-}7$$

The function v ( ) is taken as natural logarithm so that $v = \ln u$, and u is taken to have a gamma distribution. The log link leads to the linear predictor

$$\eta_{ij} = \log \mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i \qquad 5\text{-}8$$

The random effects u_i are taken to be independent distributed according to the gamma distribution with parameter λ, so that $E(u_i) = 1$ and $\mathrm{Var}(u_i) = \lambda_i$. We adopt a log-linear model for the variance of the random effect:

$$\lambda_i = \exp(\mathbf{x}'_{ij}\boldsymbol{\zeta}) \qquad 5\text{-}9$$

This model is known as the Poisson-gamma HGLM.

The log likelihood contribution of the $y|v$ part comes from Poisson density:

$$\sum_{ij} \left(y_{ij} \log \mu_{ij} - \mu_{ij} \right) \qquad \textit{(Lee et al., 2006, 180, ff)} \qquad \text{5-10}$$

and log likelihood contribution of v is

$$l(\lambda; v) = \log\ f_{\lambda}(v) = \sum_{i} \left\{ (\log u_i - u_i)/\lambda - \log(\lambda)/\lambda - \log\Gamma\left(1/\lambda\right) \right\}. \qquad \text{5-11}$$

5.2.4: Hierarchical generalized linear models with structured dispersion

Heterogeneity is common in many kinds of data and it arises from various sources. It is often associated with unequal variances. If heterogeneity is not properly modelled it can ultimately cause inefficiency and an invalid analysis. HGLMs with structured dispersion allow the dispersion parameters to have structures defined by their own set of covariates. This results in the HGLM class of joint modelling of mean and dispersion, which avoids the necessity of developing complex statistical methods on a case-by-case basis (Lee et al, 2006).

Two interlinked models for the mean and dispersion based on the observed data *y* and deviance *d* can have:

$$E(y_i) = \mu_i\ , \quad \eta_i = g(\mu_i) = \mathbf{x_i^t}\ \boldsymbol{\beta}, \quad \text{var}(y_i) = \varphi_i\ V(\mu_i) \qquad \text{5-12}$$

$$E(d_i) = \varphi_i\ , \quad \xi_i = h(\varphi_i) = \mathbf{g_i^t}\ \boldsymbol{\gamma}, \quad \text{var}(d_i) = \ 2\ \varphi_i^2 \qquad \text{5-13}$$

(Lee et al, 2006, 85, ff)

where g_i is the model matrix of explanatory variables used in the dispersion model, it is the HGLM with a gamma variance function. In the above equation the dispersion parameters are no longer constant, but can vary with mean parameters. In the GenStat software system, dispersion terms are added to the model by using the DTERMS command. This represents the variance associated with different observations that have the same value of the explanators.

5.2.5 H-likelihood

Lee and Nelder (1996) introduced the h-likelihood for inferences in HGLM. Each part of the model is evaluated by using the h-likelihood for that section. The Table 5.1 shows the likelihood which corresponds to fixed, random and dispersion part of models. The details of this are given in appendix A5.1.

Table 5.1: Likelihood used in HGLM

Part of model	**Likelihood**
Fixed part	h-likelihood for fixed part
Random part	h-likelihood for random part
Dispersion part	Adjusted profile likelihood (APL) or Extended quasi likelihood (EQL)

5.2.6 Previous research about age, gender, and mode of travel

A number of researchers have carried out various studies to identify risk factors for the age and gender groups for various modes, some of which are summarised here.

Zhang et al (2000) carried out a study in Ontario, Canada to examine factors affecting the severity of motor vehicle traffic crashes (MVTC) from 1988 to 1993 involving elderly drivers aged 65 and above. The crashes in which at least one driver was 65 or older related to automobiles or vans/light trucks were used. The dataset included 711 fatal injury crashes, 3,103 major injury, and 14,329 minor injury crashes. In this study factors of age, gender, and various other driver characteristics (normal, medical condition, use of alcohol, fell asleep etc), and environment were examined. Multivariate logistic regression was used to calculate the estimated relative risk as an odds ratio (OR) while controlling for compounding factors. It was observed that crashes involving elderly male drivers were 1.4 times as likely to be fatal as those of female elderly drivers. It was also found that failing to yield right of way / disobeying traffic signs, non-use of seat belts, intersections without traffic control, roads with a high speed limit, head-on collisions, two vehicle turning collisions, and overtaking manoeuvres were strongly related to an increased risk of fatal injury in crashes among elderly drivers. It was suggested that in order to reduce the severity of crashes involving elderly drivers, strategies should target specific factors such as head-on collisions, single vehicle

collisions, and traffic control at intersections whereas driving conditions such as medical / physical conditions and driver actions such as failing to yield right of way / disobeying traffic signs should be examined further.

Keall (1995) estimated the pedestrian risk of road accident injury in New Zealand. The estimated risk of a road accident was disaggregated by gender and age. In this study risk was estimated by dividing the number of casualties with exposure. The numbers of pedestrian casualties were extracted from the Land Transport Safety Authority Traffic Accident Report (TAR) system whereas exposure to pedestrian road accident risk was derived from the New Zealand Travel Survey. It was found that pedestrians under 10 years old and over 70 years old were more likely to be injured in a reported accident, both per road crossed and per hour of walking, than other age groups. The risk to the elderly was reconsidered in the light of the greater susceptibility to fatal injury related to age. It was found that only those over 79 years old were regarded as being at risk (2 percent of the population). It was also found that both elderly and young people spend a greater proportion of their travelling time as pedestrians than other age groups. Females spend considerably more time walking than do males. Pedestrian in their 20s cross roads more frequently per hour of walking than any other age group. Road crossing frequency was found to decline with increase in age group.

Madani and Janahi (2006) carried out a study in Bahrain to analyse pedestrian injury accidents using relevant exposure risk rates to identify the most vulnerable groups of pedestrians in terms of their personal characteristics. The characteristics investigated in this study were gender, age, nationality, and educational background. The pedestrian injury accident data files for 1995 obtained from Traffic and Licensing Directorate were used. The expected number of pedestrian accidents for gender and age groups were estimated by using an accident occurrence ratio and the proportion of population of that age group. The chi square test method was used to compare the observed accident frequencies for each category of pedestrian with the expected accidents according to their relevant proportion in the pedestrian population. It was concluded that male pedestrians have more exposure risk to accidents than females. In terms of age groups the most vulnerable were children under 12 years of age and people over 50 years of age. In terms of the nationalities there was indication that non-locals had a higher accident risk than locals whereas educated pedestrians are less likely to be involved in accidents.

Hijar et al (2000) conducted a study to identify the risk factors for motor vehicle accidents related to driver, vehicle, and environment in Mexico. The study population consisted of drivers of all motor vehicles that drove the Mexico-Cuernavaca highway from July to September of 1996. A case and control design was used. For each case driver considered, who was involved in an accident, one control driver was selected who had completed the trip on the highway without being involved in road accident. The information about the case drivers was collected by interviewing the drivers using a structured questionnaire or from passengers who were accompanying the driver in accidents where the driver died. Control drivers were selected randomly at the end points of the highway. The logistic regression was used. It was found that a higher risk was associated with drivers under 25, frequent travel, travelling to work, alcohol consumption, travel on a weekday, under adverse conditions, and in the direction of travel on the Mexico-Cuernavaca road. It was suggested that identification of these factors involved in highway traffic accidents may help in the identification of prevention measures for reducing the number of motor vehicle accidents.

Bird et al (2006) carried out a study to establish the association between land use and road traffic casualties involving non-motorised traffic. This study was carried out in Newcastle upon Tyne, in the north-east of England. The pedestrian and cyclists casualty information from 1998 to 2001 was obtained from the local government traffic accident unit while land-use data was collected using digital maps obtained from Edinburgh University's Digimap service. Log-linear models with negative binomial distribution were developed using non-motorised casualties as the response variable whilst primary functional land use, population density, and junction density were used as explanatory variables. The logarithm of length of the roads was used as offset. A total of 16 separate models were developed for each combination of cyclist and pedestrian, adults and children, working and non-working hours in the city centre and suburban analysis zones. It was concluded that during working hours, pedestrian casualties are particularly associated with retail and community land use. Priority should be given to reducing pedestrian casualties associated with retail outlets (probably shops) during working hours, and with retail outlets (almost certainly clubs and bars in city centres) during non-working hours. For cyclists' greater frequency of casualties during working hours in non-pedestrianised areas are associated with greater land-use.

Umar et al (1996) carried out a study to determine the impact of running headlights on conspicuity-related motorcycle accidents in Malaysia. The Generalized linear model with Poisson distribution and log link was used to describe the frequency of conspicuity-related

motorcycle accidents. The explanatory variables used consisted of: influence of time trends, changes in recording system, effect of fasting during month of Ramazan, and Balik Kampong which is a religious holiday unique to the multicultural society of Malaysia. In order to overcome the over-dispersion of data, the quasi-likelihood technique was used. From the modelling results it was concluded that time is a positively significant variable with an increase of 0.5 percent conspicuity-related accidents per week. The new recording system improved the quality and quantity of data. An increase of 40 percent in conspicuity-related motorcycle accidents was observed after the introduction of the new system. It was also found that number of accidents increased by 41 percent in the fasting season for which changes in travelling and social religious activities were a possible cause. The Balik Kampong variable was found to be non-significant. It was also shown that the use of running headlights reduced conspicuity-related accidents in Malaysia by 29 percent.

Legge et al (1998) studied age and gender differences in the rates of crash involvement of Western Australian drivers. The Road Injury Database of the Road Accident Prevention Research Unit from January 1989 to December 1992 was used. The population examined was all drivers of cars, station wagons, and related vehicles involved in damage-only, injury and fatal crashes. Risk ratios were estimated for various age groups. It was found that drivers under 25 years of age were involved in 35 percent of crashes, compared to 3 percent for drivers aged 70 years and over. Drivers aged under 25 had the highest rates based on both a population and a licence basis, but after taking distance travelled into consideration the crash involvement of both groups were almost same. Females had higher rate of crash involvement than males in all age groups. It was also found that the youngest groups of drivers had proportionately more single vehicle crashes, drivers aged 30 to 59 had more same-direction crashes and drivers over 60 years, particularly over 75 years, had more direct and indirect right angle crashes. It was concluded that the risk of crashes varies according to ability, experience, and psychological function, which are related to age.

Fontaine and Gourlet (1997) examined the reports of fatal pedestrian accidents in France to improve the understanding of these accidents and to propose some suitable action. A total of 1,289 fatal pedestrian accidents which occurred from March 1990 to February 1991 were considered. The age, gender, movements, change of mode, and alcohol impairment characteristics were analysed. The accidents were classified into four categories. It was found that elderly pedestrians crossing the road in an urban area at a junction (often controlled by

traffic lights) composed of 42 of all fatally injured pedestrians. These accidents occurred on weekdays between 7 a.m. and noon, or between 2 p.m. and 6 p.m. A second category making up 34 percent of pedestrian fatalities was those with high alcohol concentrations involved in night-time accidents. Most of these accidents took place at night time, on weekends and not at a junction. A third category of children running or playing made up 13 percent of all fatally-injured pedestrians. A fourth category included secondary accidents, change of transport mode and consisted of 11 percent of total fatally-injured pedestrians. It was suggested that information campaigns and lifelong safety education programmes for pedestrians could be considered to stress the particular dangers faced by them.

The literature review in this section highlights the importance of identifying the high risk groups that could be used by planning organisations for improving the road safety. In most of these studies the particular emphasis is given on identifying target groups which could be used to elevate risk awareness and ultimately to improve road safety. Risk ratios for different age and gender groups were highlighted. It was also found that different measures of exposure were used by various researchers based on the availability of data. Bird et al (2006) used the road length as exposure through offset variable, Madani and Janahi (2006) used population, while Keall (1995) used an estimate of pedestrian time spent in walking.

Legge et al (1998) found that drivers aged under 25 years had higher rates of accident involvement per person-year than drivers aged 70 years or over, but after taking distance travelled into consideration the accident involvement of both the age groups was same. This shows that risk ratios will vary depending on the exposure considered (i.e. population, distance travelled, number of licence holders).

In the present study, the main focus is given to identifying the risk values for different age group, gender and vehicle type on a national scale, which could be used by various planning and road safety agencies to improve road safety.

5.3 DATA USED

Three data sources were used for the present study. The numbers of casualties were extracted from STATS 19 data for the years 2001-5. For each of these years the distance travelled by

different age groups was extracted from NTS data and population numbers were extracted from National Statistics, United Kingdom. All of these are described in detail below.

5.3.1 Combined road accidents and casualty data (STATS 19)

The STATS 19 road accident statistics of Great Britain from 2001 to 2005 are used for the present study. The year-wise accident and casualty information of STATS 19 data were joined together in MS Access in order to extract the number of casualties disaggregated by day of week, month, year, age group, gender, and mode. MS Access queries were used to create two new fields of vehicle class and age group, which are shown in Table 5.2 and 5.3; this ensured compatibility between the categories used in different datasets. After this, these files were exported to SPSS to develop a new dataset consisting the information of all the road casualties that occurred from 1st January 2001 to 31st December 2005. Five different datasets each representing a single mode of the new classification were developed, each with a 29,216 records. Car, walk, bicycle, motorcycle, and bus modes were considered in this study. This was mainly done due to the limitations of the GenStat software which was unable to accommodate a large amount of data, such as the whole dataset of all records for all modes, in estimating the HGLMs.

Table 5.2: Reclassification of the modes considered

S.No	Vehicles classified in STATS 19	New Classification	S.No	Vehicles classified in STATS 19	New Classification
1	Pedestrian	Pedestrian	6	Taxi	
2	Pedal Cyclist	Pedal Cyclists	7	Car	Car
3	Moped		8	Bus or Coach	Bus
4	Motorcycle (up to 125 cc)	Motor Cyclists			
5	Motorcycle (over 125 cc)				

5.3.2 National travel survey data (NTS Data)

The distance travelled per person-year by gender, mode, and age group was obtained from the DfT. The distance travelled was given in miles for each age category by walk, bicycle, car driver, car passenger, motorcycle, and local bus. The car driver, car passenger, and taxi were

added together to obtain the distance travelled by car. This does not include taxis running only with drivers because the DfT does not collect such information. The distance travelled by bus is the distance travelled in local buses, which excludes intercity buses. The age groups considered for distance travelled by the DfT are shown in Table 5.3. The National Travel Survey (NTS) provides information about personal travel within Great Britain and it also monitors the trends in travel behaviour. The Ministry of Transport commissioned the first NTS in 1965/66 which was repeated in 1972/1973, 1975/1976, 1978/1979, and 1985/1986. From 1988 the NTS became a continuous survey and fieldwork is conducted on a monthly basis. The NTS involved posting contact letters, making initial contact, arranging interviews, providing the travel diaries, making a reminder call, mid-week check call, conducting the pick-up interview at the end of travel week, and transmission of the data. During the process the information about the seven-day travel record, long-distance journeys, fuel and mileage chart are recorded. After the collection and brief checking of the seven-day travel diaries, the information is entered into the Diary Entry System (DES). The data is then delivered to the DfT after making several checks and verification about the cleanness of the data.

5.3.3 Population data (2001-2005)

The population of Great Britain from 2001 to 2005 was obtained from the annual abstracts of statistics produced by Office for National Statistics, United Kingdom. The data was available separately for England, Scotland, and Wales. The age categories in data available from the Office for National Statistics were not the same as in the distance travelled data which was provided by the DfT. Consequently, the population age group data was rearranged to match the age classification of the distance travelled data. In this rearrangement of the population data, it was supposed that total yearly population of males and females was uniform within each of the ranges. The age groups considered are shown in Table 5.3.

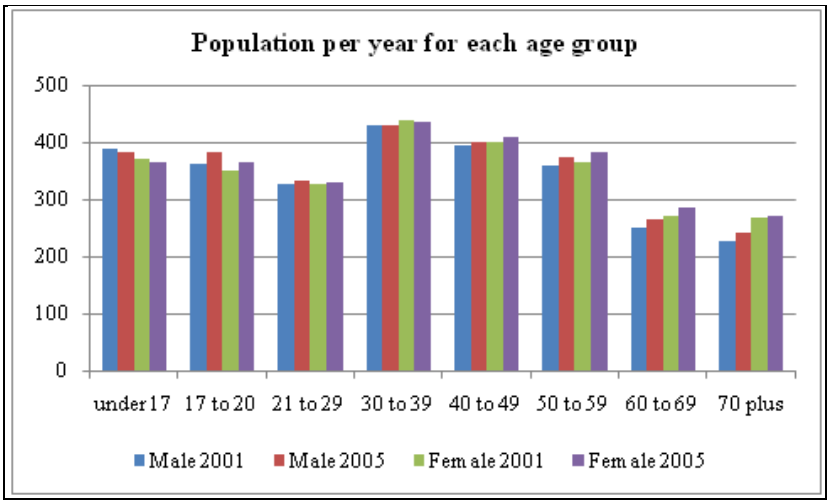
Table 5.3: Age groups considered for the present study

Age Band	1	2	3	4	5	6	7	8
Age group	Under 17	17 to 20	21 to 29	30 to 39	40 to 49	50 to 59	60 to 69	70 plus

The details of population per year in each age group are shown in Figure 5.1 from which it is found that:

- The population of Great Britain was 57.42 million in 2001 which increased to 58.41 million in 2005.
- Males made up 49 percent of the total population and females 51 percent.
- The 30 to 39 age group had a higher population per year than other age groups, followed by the 40 to 49 age group.
- The 60 to 69 and 70 plus age groups had a lower population per year than the other age groups.
- The number of persons per year in the age group under 17 is on decline.

Figure 5.1: Population per year of each age group (in thousands)



Source of data: Office for national statistics, UK (2011)

5.4 DATA ANALYSIS

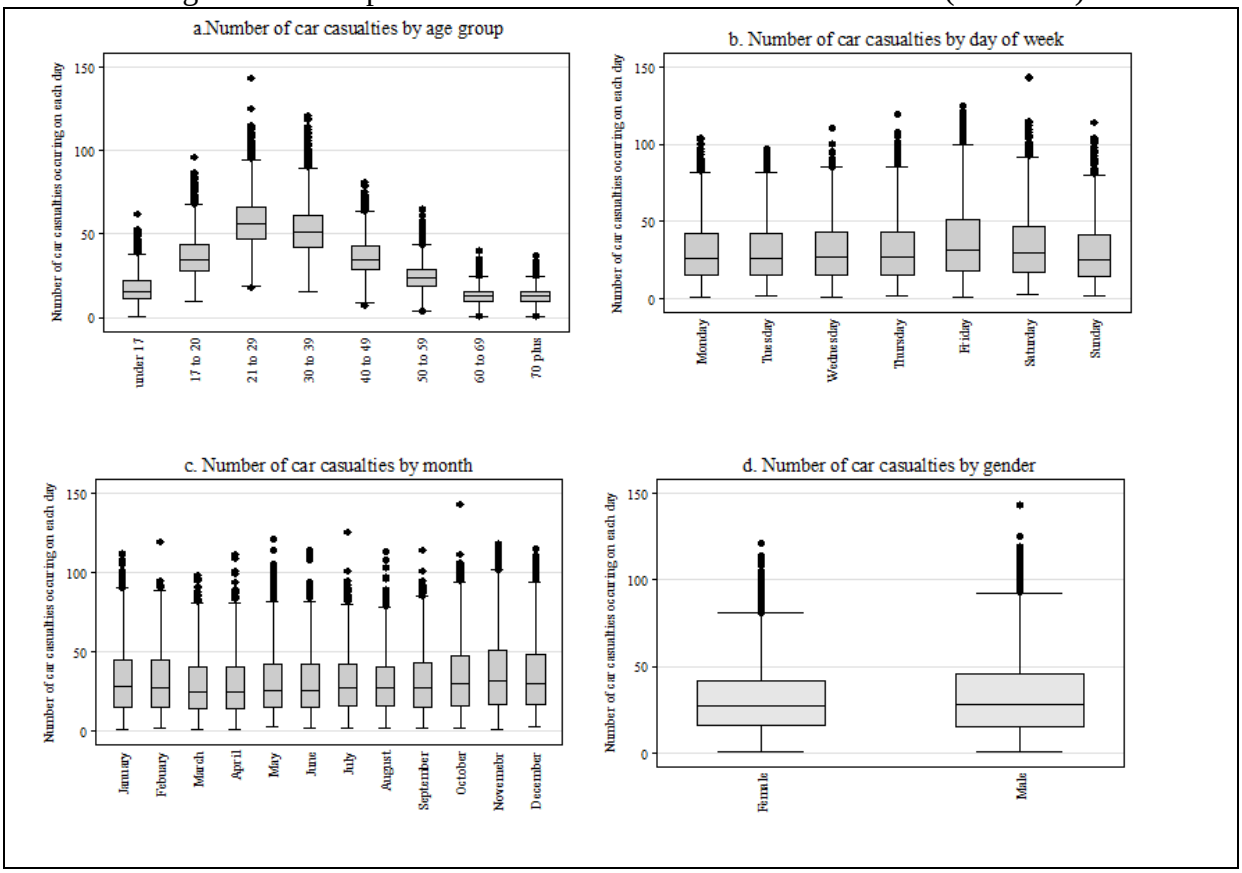
STATS 19 data and travel data used in this study are analysed below:

5.4.1 STATS 19 data (2001-2005)

Five new datasets were developed by combining the accident and casualty information of STATS 19 data from 2001 to 2005, each representing a mode. MS Access and SPSS were used to extract the number of casualties' information disaggregated by age group, gender, and mode. The box plot for the casualty data are shown in Figure 5.2 to 5.6 and each dataset is analysed as follows:

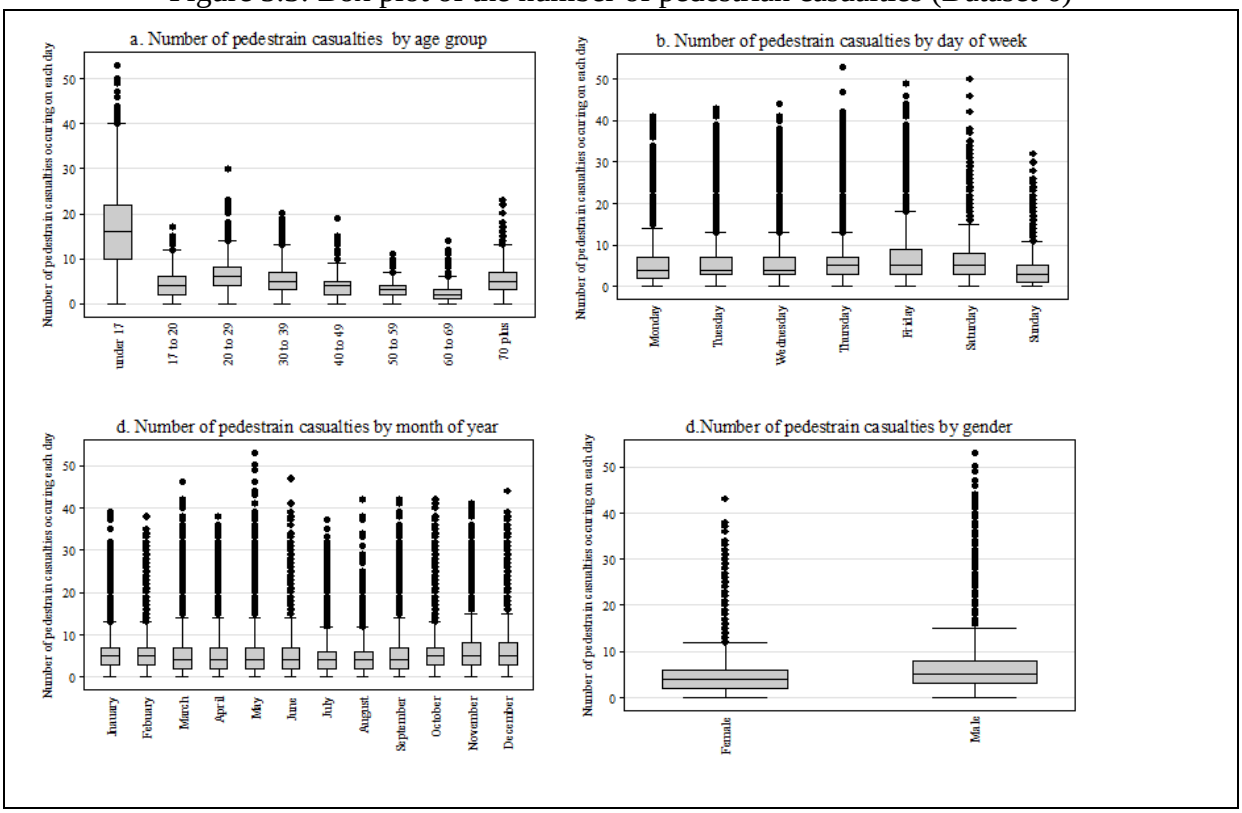
- The age group under 17 have a higher number of casualties for walking and cycling modes while the age groups 30 to 39 and 21 to 29 were respectively involved in more motorcycle and car casualties than any other age group.
- Elderly (70 plus) pedestrians and bus users have higher casualties than most age groups.
- The casualties for each of mode decreases with increasing age after a certain age group as a result of being mature and experience.
- A difference in the number of casualties existed between weekdays and weekends (especially Sunday) across all modes. Saturday had slightly higher pedestrian and car casualties than the first three weekdays.
- Summer months had higher cyclist and motorcyclist casualties while car users had higher casualties in winter months.
- A comparatively small difference in casualty numbers was observed between male and female car users in comparison to other modes where a higher number of casualties were male.

Figure 5.2: Box plot of the number of casualties for car users (Dataset 5)



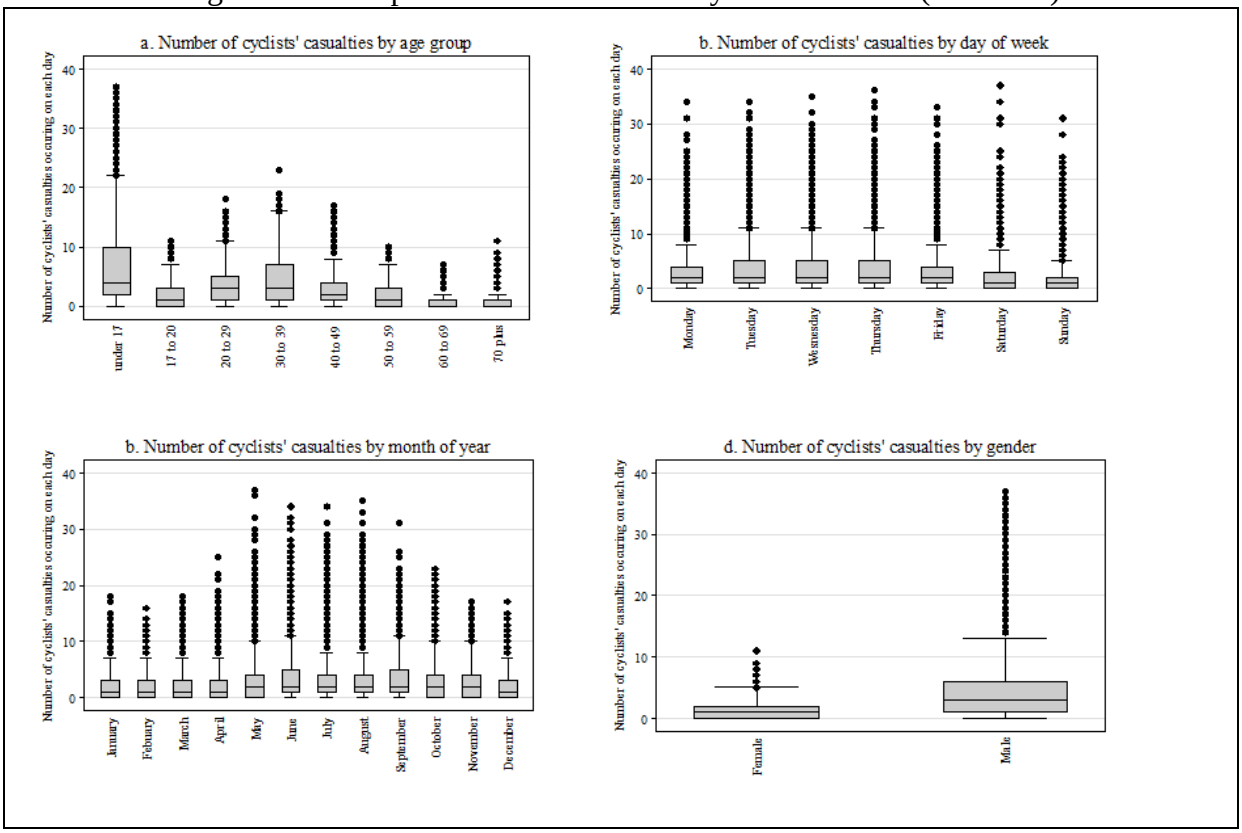
Source of data: Department for Transport (2011)

Figure 5.3: Box plot of the number of pedestrian casualties (Dataset 6)



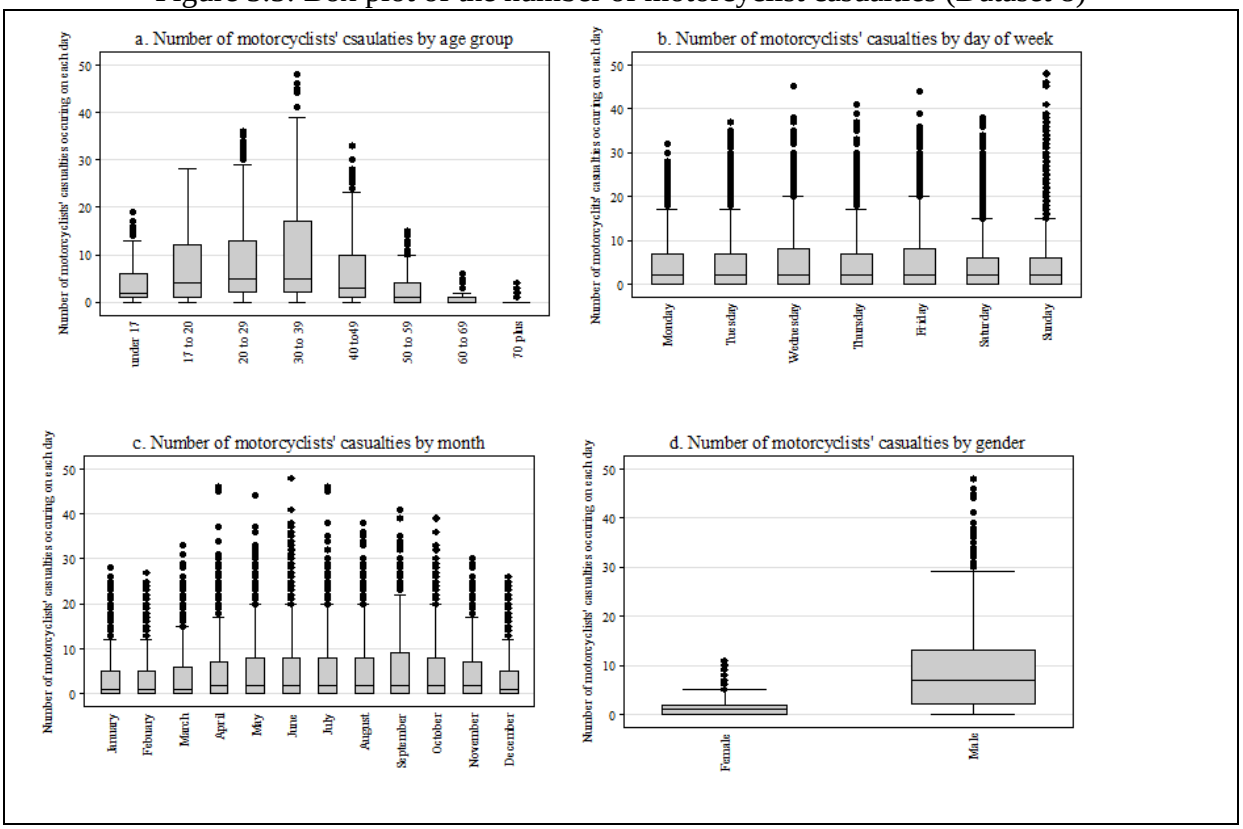
Source of data: Department for Transport (2011)

Figure 5.4: Box plot of the number of bicyclist casualties (Dataset 7)



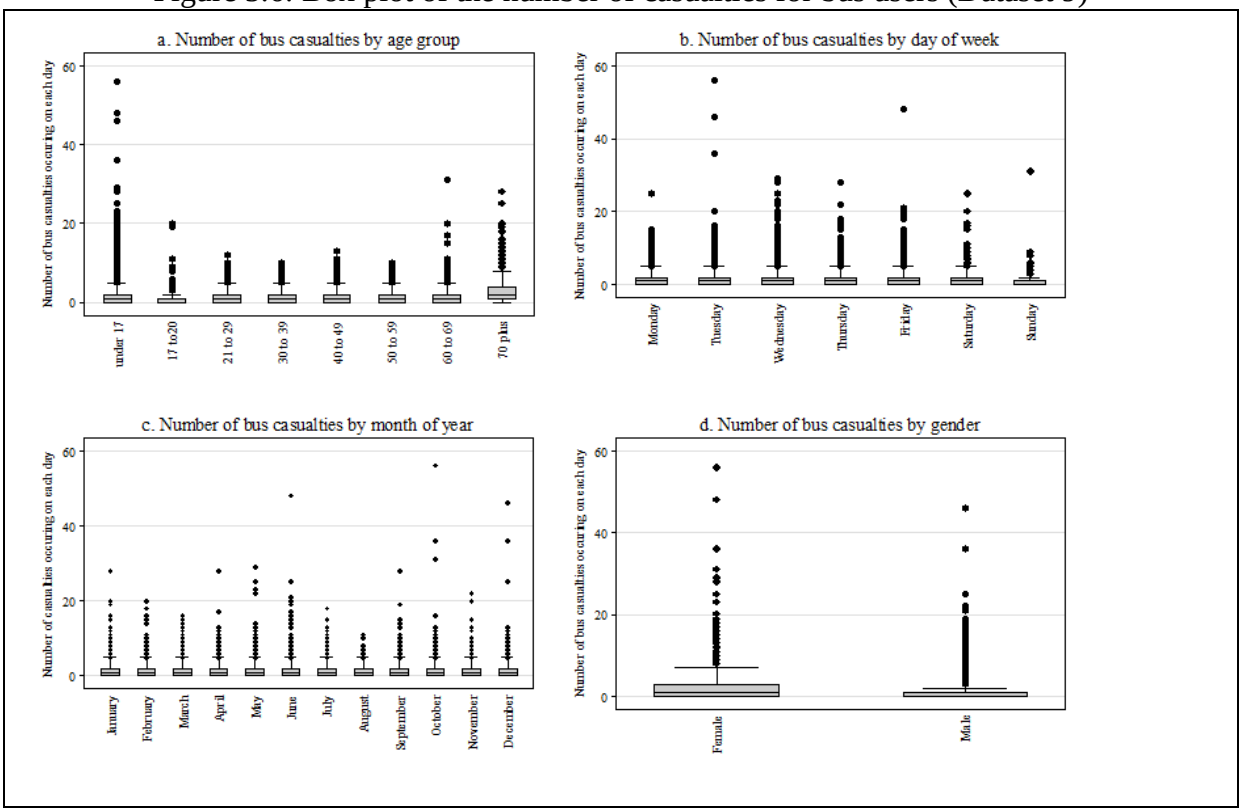
Source of data: Department for Transport (2011)

Figure 5.5: Box plot of the number of motorcyclist casualties (Dataset 8)



Source of data: Department for Transport (2011)

Figure 5.6: Box plot of the number of casualties for bus users (Dataset 9)



Source of data: Department for Transport (2011)

5.4.2 Travel data (2001-2005)

The annual distance travelled per person from 2001 to 2005, disaggregated by age and gender, extracted from the NTS data was obtained from the DfT. The data for the years 2001 to 2005 is shown in Figure 5.7 and is analysed below:

- 17 to 20-year olds walked more than all other age groups. Females in the age groups 21 to 39 walked more than males in the same age groups. Older males walked more than females.
- Males cycle more than females. A singular peak in 2001 was observed for males of 17 to 20 years of age. The distance travelled by males cyclists between 21 and 49 years old increased from 2001 to 2005. Cycling by females aged 21 to 39 years and 70+ decreased in 2005 in comparison to 2001 whereas for all other age groups it increased.
- Males of all age groups travel more by motorcycle than do females. A higher distance was travelled by 40 to 49 year olds in 2005. Older people travel less by motorcycle, with people over 70 travelling less by motorcycle than any other age group.
- Males travel a greater distance by car than females. The distance travelled per person increases with age until 50, after which it decreases. The highest distance per person per day was travelled by the 40 to 49 age group. Males from 17 to 59 years of age travelled less distance in 2005 than in 2001. Females other than those between 17 and 29 travelled more in 2005 than in 2001. This was particularly so for females aged 40 to 49.
- Young persons of age between 17 to 20 years travelled more by bus than all other age groups. A huge difference was observed in comparison to other age groups. Females above 40 travel more on buses than do males. The distance travelled by males over 60 years old was slightly less in 2005.

Figure 5.7: Graph showing distance travelled per person (kilometres) for different modes

Source of data: Department for Transport (2011)

5.5 MODEL DEVELOPMENT

The first step in the model development was to identify which explanatory variables would be considered for use as random effects. The interaction between month and year variables was selected to be the random part as theory suggested that it would be an appropriate for this: the yearly instance of each month was considered to be a sample of larger population whereas all other variables had fixed categories and they could not so readily be viewed as sample of a larger population. Because month is also included in the fixed model, this represents the concept that number of casualties occurring in each month of the year will follow a general trend (the fixed effect) but this will also vary between years (the random effect). After this,

the fixed part was identified by stepwise inclusion of variables. Variables of age group, gender, interaction of age group and gender, day of week, month, time, holidays, New-Year and Christmas holidays were used in the fixed part. The total distance travelled was not considered in the full model as an explanatory variable as it was subsumed by age group due to its synthesis. The h-likelihood for the fixed part was monitored through the model development. After selecting the fixed part, the random part was reviewed. During this, the h-likelihood for the random part was monitored. After this, dispersion terms were identified and were included one by one into the model.

Lee, Nelder and Pawitan (2006, p158) recommend that when dispersion terms are added in the model, the adjusted profile likelihood (APL) is an appropriate measure of model performance. However, this measure was found to be unreliable in the models developed here: in the models of bicycle, motorcycle and bus casualty data the APL did not always improve when a further variable was added to the dispersion part. Due to this, the extended quasi likelihood (EQL) was adopted instead to compare the performance of dispersion terms in the models: this measure was found to be satisfactory. The logarithm value of yearly population of age group was preferred for use as an offset, which allowed for the variation in population for age group, gender and year. This yields a model of casualty rate per person-year. Five models, each representing a mode, were developed and variables were removed from each model in steps. The h-likelihood was monitored as shown in Table 5.4.

Table 5.4: Model development sequence and likelihood used

Step	**Model**	**Model development sequence**
0	Random part	Month.Year
1	Fixed part	h-likelihood for fixed part
2	Random part	h-likelihood for random part
3	Dispersion part	Extended quasi likelihood (EQL)

5.5.1 Variables used

The following variables were used in the models.

1. Logarithm (Population disaggregated by age and gender)
2. Age group (in years) (8 levels)

   <17 ｜17-20｜21-29｜30-39｜40-49｜50-59｜60-69｜70 plus
3. Gender (2 levels)
4. Interaction of age group.gender (16 levels)
5. Day of week (7 levels)
6. Month (12 levels)
7. Time as variate (values in days from 1 to 1826; 1st January 2001 to 31st December 2005)
8. Public holiday
9. New-Year holiday
10. Christmas holiday

5.5.2 Basic Model structure

In this chapter all models were developed as shown below and then each variable was removed from the model and its effect on the h-likelihood was monitored. An offset variable was also introduced to represent the exposure. Population, total daily distance travelled and yearly distance travelled were tested in offset to identify the most suitable one. It was found that population performed better than the other two variables in h-likelihood results (see appendix Table A5.2): this leads to models that can be interpreted in terms of casualty rate per person-year stratified by age group and gender. The hierarchical generalized linear model with Poisson-gamma log link and generalized estimation equation (GEE) with autoregressive error structure (AR1) is used which is described below:

HIERARCHICAL GENERALIZED LINEAR MODEL (HGLM)

Let Y be the number of casualties occurring on each day disaggregated by age, gender, and vehicle class.

OFFSET: Logarithm (Population of age group by gender)

Fixed effect: Age group + Gender + Age group. gender + Day of week + Month + Time + Holiday + New-Year + Christmas

Random effects: Month.Year

Dispersion Terms: Age, Gender, Month, Day of week

Link: Logarithm

Errors: Poisson-gamma

GENERALIZED ESTIMATION EQUATION (GEE)

A generalized estimation equation model with negative binomial regression having AR1 errors was also developed consisting of the following variables. The results were then compared with the preferred model developed by using HGLM.

OFFSET: Logarithm (Population of age group by gender)

Fixed effects: Age group + Gender + Age group. gender + Day of week+ Month + Time + Holiday + New-Year + Christmas

Link: Logarithm

Errors: Negative binomial with autoregressive (AR1) error structure

The following variables were considered and tested in the offset as measures of exposure in the models.

1. Logarithm (Population): This variable represented the population of each year disaggregated by age group and gender.
2. Logarithm (Total daily distance travelled): This variable represented the total distance travelled each day by all members of the population in the specified age group and gender. The daily distance travelled per person was multiplied by the Population of age and gender group.
3. Logarithm (Yearly distance travelled per person): This variable represented the distance travelled each year per person by age group and gender.

A full model as shown above with each of these variables as an offset for Dataset 5 (Car mode) was developed and its h-likelihood values were compared. From Table A5.2 shown in the Appendix it was found that the model with Logarithm of Population as an offset had better h-likelihood than the other two models (A2 and A3) so it was preferred. The h-likelihood of the model with population as an offset (model A1) was better by the value of 1,097 and 978 from the model A2 and A3 with total daily distance travelled and yearly distance travelled per person respectively. The population was preferred as the offset according to the goodness of fit: it offers the advantage that the model coefficients can be interpreted in terms of the casualty rate per person-year. By contrast, use of distance travelled as offset assumes implicitly the uniformity of distance travelled over days of the year because no suitable correction factors for different modes, age and gender groups were available. Due to these reasons, population was preferred as offset in the model.

5.6 MODEL SELECTION PROCESS, GOODNESS OF FIT AND MODEL CHECKS

Following sections shows the results of the models developed by using each of five datasets which represents the modes.

- Car (Dataset 5)
- Walk (Dataset 6)
- Bicycle (Dataset 7)
- Motorcycle (Dataset 8)
- Bus (Dataset 9)

The following model selection procedure was applied to select the most appropriate model to estimate the number of casualties occurring on each day for age and gender groups.

5.6.1. Model selection process, goodness of fit and model checks for Dataset 5 (Car)

As in chapter 4 it was found that car is involved in greater number of road accidents than other modes, due to which in this chapter the investigation started by identifying the relationships of age groups and gender in this case. This section shows results of the models developed, goodness of fit of the preferred model and various checks to validate the model.

5.6.1.1 Hierarchical generalized linear model (HGLM) Car:

The Hierarchical Generalized Linear Model (HGLM) with Poisson-gamma distribution with log link was used. In the first step, the full model as shown in section 5.5.2 was developed. The h-likelihood values obtained for the fixed, random and dispersion parts are shown in Table 5.5. After this, individual variables were removed from the model to investigate their partial effect on the h-likelihood. In the first step variables of age.gender, day of the week, month, time, holidays, New-Year holidays and Christmas holidays were removed in turn and the h-likelihood of the fixed part was compared. This confirmed that the full model had better h-likelihood for the fixed part of the model, and that removal of any variable would result in substantially reduced preference. It was found that day of the week had the highest effect among the listed variables as its removal from the fixed part reduced the h-likelihood by 1,840 with only 6 degrees of freedom. Out of all the variables, Christmas holidays have the least effect of a change of 62 in h-likelihood, which is statistically significant at the 5 percent level according to the likelihood ratio test.

In the second step, the month.year interaction was removed from the random part of the full model, keeping the same fixed and dispersion parts. It was observed that removal of month.year reduced the h-likelihood of fixed part by 234 with 1 degree of freedom (corresponding to the variance of the random effect) confirming that it contributes substantially to model performance. In the third step, the variables were removed in turn from the dispersion part of the model while keeping the same fixed and random parts. The results showed that age group is the most important variable which affected the extended quasi likelihood (EQL) by 1,164 with 7 degrees of freedom. The removal of Gender reduced the

EQL by only 3 with 1 degree of freedom. This suggests that Gender is equally variable. This is also evident from the coefficient obtained for Gender in the dispersion model which is found to be non-significant. The removal of other variables also reduced EQL, details of which are shown in Table 5.5. It was concluded that the full model had significantly better h-likelihood results than simplified models.

In the full model, the different variances are represented by the coefficients of λ and ϕ. The exponential of the coefficient λ which represents the variance of the random effects, in the present case it is found to be -6.73 (exponential is equal to 0.0011) and is significantly different from 0 with a *t* value of -30.72. On the other hand ϕ represents the variance of individual observations as used in regression analysis, though the dispersion model allows for this to vary according to the variables. In this case, the random component of the month.year compares each month of the year to the usual value for that month. The month in dispersion part quantifies the variation present in a particular month: for example observations in December were found to be more variable than those in other months. Detailed results of the coefficients of fixed, random and dispersion parts are discussed in the next section.

Table 5.5: Results of the h-likelihood (Dataset 5: Car)

Models	**Variables**	**d.f.**	**H likelihood and change in its value**		
			Fixed	**Random**	**Dispersion**
FIXED	**Full Model**		**201,681**	**201,257**	**201,708**
	- Age. Gender	7	+1,664		
	- Day of week	6	+1,840		
	- Month	11	+111		
	- Time	1	+72		
	- Holiday	1	+86		
	- New Year	1	+151		
	- Christmas	1	+62		
RANDOM	- Month.Year	1	**+234**		
DTERMS	- Age	7			+1,164
	- Gender	1			+3
	- Month	11			+113
	- Day of week	6			+33

5.6.1.2 Analysing the temporal effects

The procedure presented in section 2.6.2.2 was used to investigate for the presence of further substantial temporal effect that was not represented in the models. This process was carried out for only full model (shown in section 5.5.2). For this, square of time variable was added to the full model as time was already present in the linear predictor. After this improvement in the h-likelihood, coefficients and *t* values of time and square of time, and their variance inflation factors were examined.

It is observed from the results shown in appendix Table A5.3 that after adding square of time into the full model, there is no improvement in the h-likelihood. The estimated coefficient of the square of time was also found to be non-significant whilst time and square of time variables had high variance inflation factors (value of 16). This shows that no substantial temporal effect remains in the model that can be represented by the quadratic terms.

5.6.1.3 Split sample tests

In order to check the consistency of the model parameters, split sample validation tests were carried out. To do this, the dataset was randomly partitioned into two, each with 14,608 observations. The datasets A, B and C were used to check and validate the results of the full model by comparing the coefficients of all the three models and observing their h-likelihoods.

GenStat software was used to estimate the model parameters of Datasets B and C which were then compared. The results in Table 5.6 show that h-likelihood for Dataset B was slightly better than for Dataset C. The h-likelihood for the fixed part was better by a value of 255, random part by a value of 263 and dispersion part by a value of 250.

Table 5.6: h-likelihood results of the split sample (Dataset 5: Car)

Model	No of observations	h-likelihood		
		Fixed Part	Random Part	Dispersion Part
Dataset A	29,216	201,681	201,257	201,991
Dataset B	14,608	100,706	100,313	101,000
Dataset C	14,608	100,961	100,576	101,250

After this the coefficients of the fixed part were compared between models A, B and C. The *T* test was used to compare the coefficients of Datasets B and C. T_{BC} values were estimated by using the formula 2-32. It was found from *T* test values that all of the coefficients (except age group 30-39, Christmas holidays and constant) of model B are not significantly different from the coefficients of model C as the estimated values of T_{BC} are less than 1.96. However, the change in the constant was of less concern as it represented the group mean which can vary in dataset B and C. It is also to note that out of 37 variables only 3 were found to have changed significantly. The comparison of coefficients and *t* values are shown in Figure 5.8 and Table 5.7. In summary:

- The coefficient of age group had significant *t* values and expected signs in all three models.

- The coefficient of Gender was positive and had significant *t* values in all three models. The coefficient of 30 to 39.male had a significant *t* value in models A and B but was non-significant in model C. The coefficient of 60 to 69.male had a significant *t* value in model C only.

- All coefficients of Day of week differed from each other and had significant *t* values in all three models except Thursday which was non-significant in all three models.

- All coefficients of months differed from each other and had significant *t* values in all three models except January, February, July, August, and September. Except August all these months had non-significant *t* values in all three models while August had significant *t* values in model B only.

- The coefficients of time, holidays, New-Year holidays and Christmas holidays had similar sign and significant *t* values in all three models.

Figure 5.8: Comparison of coefficients of full model HGLM for coefficient validation (Car)

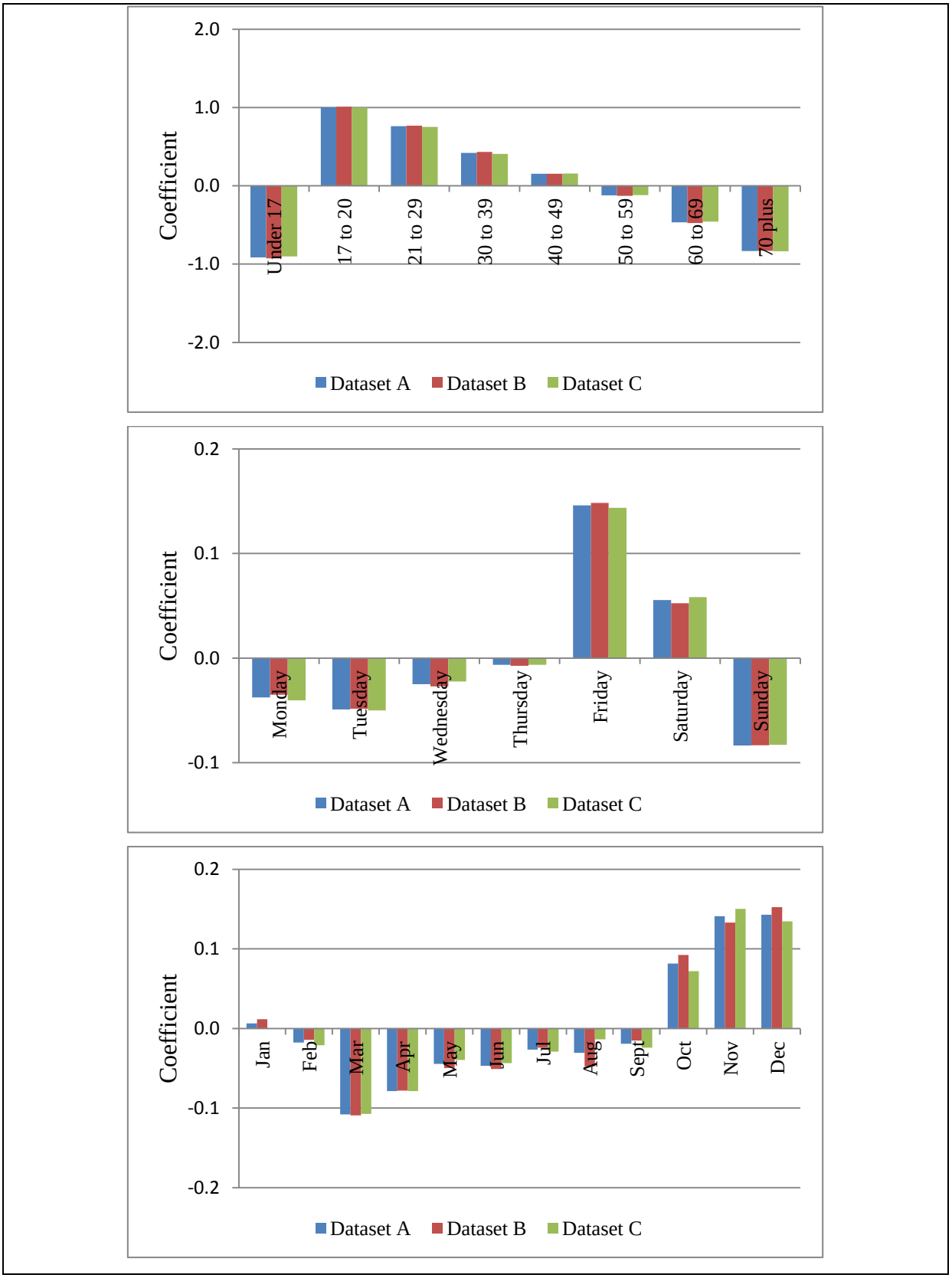


Table 5.7: Comparison of coefficients and t values of full model HGLM (Split sample Data)

Variables	Comparison of the coefficients and *t* values of the Models						
	Model A		Model B		Model C		
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	T_{BC}
Under 17	-0.913	-100.18	-0.926	-72.04	-0.900	-69.61	1.425
17-20	1.003	151.92	1.008	107.56	0.999	107.39	-0.696
30-39	0.420	79.70	0.432	58.48	0.407	54.18	-2.341
40-49	0.155	27.59	0.153	19.72	0.156	19.10	0.236
50-59	-0.123	-20.81	-0.128	-15.59	-0.120	-13.94	0.686
60-69	-0.467	-66.39	-0.477	-48.33	-0.457	-45.48	1.416
70 plus	-0.834	-115.37	-0.831	-83.3	-0.837	-80.34	-0.388
Gender	0.023	13.11	0.024	9.63	0.023	9.03	-0.405
Under 17.Male	-0.280	-20.89	-0.273	-14.15	-0.287	-15.44	-0.537
17-20. Male	0.214	24.46	0.209	16.87	0.218	17.58	0.536
30-39. Male	-0.016	-2.17	-0.026	-2.55	*-0.005*	*-0.48*	1.434
40-49. Male	-0.075	-9.43	-0.065	-5.79	*-0.085*	*-7.36*	-1.251
50-59. Male	-0.131	-15.37	-0.120	-10.06	-0.142	-11.65	-1.320
60-69. Male	*-0.013*	*-1.34*	*0.002*	*0.15*	-0.029	-2.05	-1.549
70 plus. Male	0.240	23.22	0.227	15.68	0.255	17.27	1.351
Monday	-0.038	-9.15	-0.035	-6.05	-0.040	-6.90	-0.649
Tuesday	-0.049	-12.50	-0.048	-8.92	-0.050	-8.80	-0.234
Wednesday	-0.025	-6.34	-0.027	-4.88	-0.022	-4.00	0.570
Thursday	*-0.006*	*-1.66*	*-0.007*	*-1.34*	*-0.006*	*-1.14*	0.113
Saturday	0.056	14.34	0.052	9.71	0.058	10.47	0.779
Sunday	-0.084	-19.69	-0.083	-13.63	-0.083	-14.03	0.049
January	*0.006*	*0.40*	*0.012*	*0.76*	*0.000*	*-0.02*	-0.499
February	*-0.018*	*-1.12*	*-0.014*	*-0.94*	*-0.021*	*-1.18*	-0.291
March	-0.108	-6.87	-0.109	-7.24	-0.107	-6.02	0.073
April	-0.079	-4.99	-0.078	-5.14	-0.079	-4.39	-0.021
May	-0.045	-2.84	-0.050	-3.29	-0.040	-2.24	0.438
June	-0.047	-2.97	-0.051	-3.36	-0.044	-2.42	0.314
July	*-0.027*	*-1.70*	*-0.024*	*-1.6*	*-0.029*	*-1.63*	-0.201
August	*-0.031*	*-1.95*	-0.048	-3.14	*-0.014*	*-0.76*	1.466
September	*-0.019*	*-1.21*	*-0.015*	*-1.01*	*-0.024*	*-1.35*	-0.377
October	0.082	5.18	0.092	6.06	0.072	4.04	-0.863
December	0.143	9.02	0.152	9.94	0.135	7.49	-0.753
Time	-0.0001	-10.57	-0.0001	-10.71	-0.00009	-9.73	-0.505
Holidays	-0.057	-9.19	-0.057	-6.54	-0.056	-6.35	0.093
New-Year	-0.142	-7.61	-0.145	-5.58	-0.138	-5.16	0.175
Christmas	-0.196	-11.85	-0.247	-10.38	-0.149	-6.50	2.970
Constant	-12.04	-464.97	-12.09	-343.16	-11.987	-336.17	2.207

Italic shows that these variables are not significant at 5 percent level.

5.6.1.4 Comparison of Coefficients (Car: HGLM and GEE-AR1)

a. Fixed Part:

The coefficients of HGLM with Poisson-gamma distribution and log link, and GEE-AR1 with negative binomial were also compared. Because the coefficients of these two models (HGLM and GEE) are estimated by using the same data, they are not mutually independent so it is not possible to test for differences between them. It is observed that HGLM cannot accommodate time series error structure (AR1) and GEE has no feature to take account of structured variance. Neither is refinement of the other. Due to this, we will look into the similarities in the estimated coefficients (similar signs and values of estimated coefficients) and what are differences in the estimated coefficients.

It was found that all coefficients that were significant at the 95 percent level in both models had the same sign, though a slight change was observed in the *t* values of the variables. Age group and interaction variables had better *t* values in HGLM except for males under 17. All month variables had better *t* values with GEE-AR1 whereas, for day of week, only Sunday had better *t* values in HGLM. The coefficients of February, July, August and September which were found to be significant in GEE-AR1 became non-significant in the HGLM model. In the same way, the coefficient of the male age group 30 to 39 was not significant in GEE-AR1 but was significant with HGLM. Although some variables changed from significant to non-significant but no variable changed its sign from one model to the other.

The age group 17 to 20 had a higher coefficient which showed the greatest casualty rate per person-year for this age group in car casualties. With increase in age, coefficient of age decreases which highlights that the casualty rate per person-year decreases with increase in age due to maturity or by getting more experience. Under 17 and 70 plus had the lowest rate per person-years in car casualty data. Gender had a positive coefficient showing greater casualty rate per person-years for males. However, the interaction of age and gender shows that females in all age groups (under 17 and 40 to 69) have greater casualty rate per person-years than the males of same age groups after allowing for their main effects.

March, April, May, June, October and December had significant *t* values. December had greater coefficient followed by November and October which shows that last three months of

year (October, November and December) are associated with greater casualty rate per person than other months. Friday has the greatest casualty rate among days of the week whereas Sunday had the least rate per person. Saturday had greater casualty rate per person than all other days of the week except (Friday). It was also found that rate per person-years for car casualties is also decreasing over time at about 3.5 percent each year as coefficient of time was found to be (-0.000096) with significant *t* value. Christmas holidays were found to have the most negative coefficient showing lower rate per person than New-Year and other Public holidays. The comparison of the coefficients for the fixed part of HGLM and GEE-AR1 are shown in Figure 5.9 and Table 5.8.

b. Random Part:

The interaction of month.year was used in the random part of the HGLM model. The parameter λ represents the variation between corresponding months in different years. It was found that λ had a coefficient of -6.73 with a significant *t* value of -30.62. The exponential of the λ represents the variance of the random component. Only 13 out of the 60 combinations of month.year interactions had significant *t* values which showed that these months in the mentioned years were significantly different from zero. The interpretation of random term can be made that March 2003 was different and had fewer car casualties than a usual March. In the same way March 2004 was different but it had higher car casualties than a usual March as it has a positive coefficient. The detailed results of significant coefficients of the random part are shown in Table 5.9.

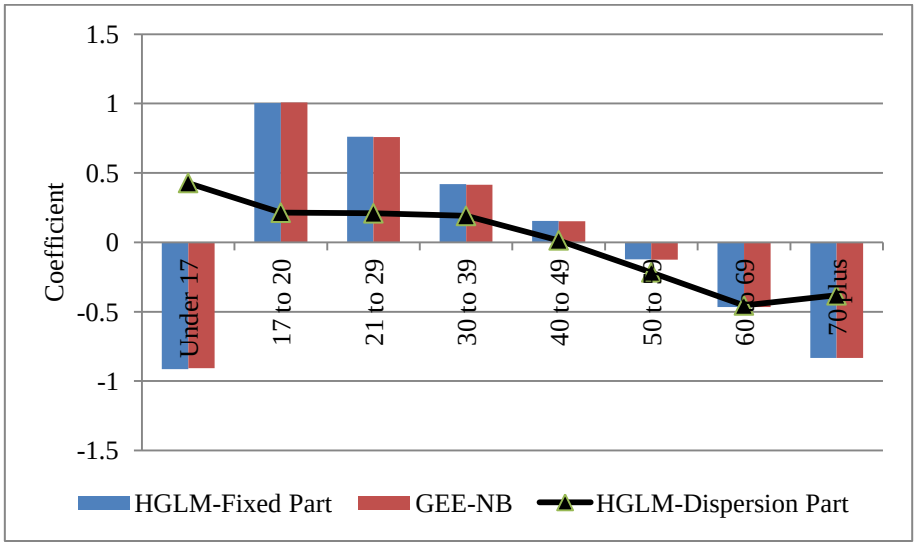
c. Dispersion Part:

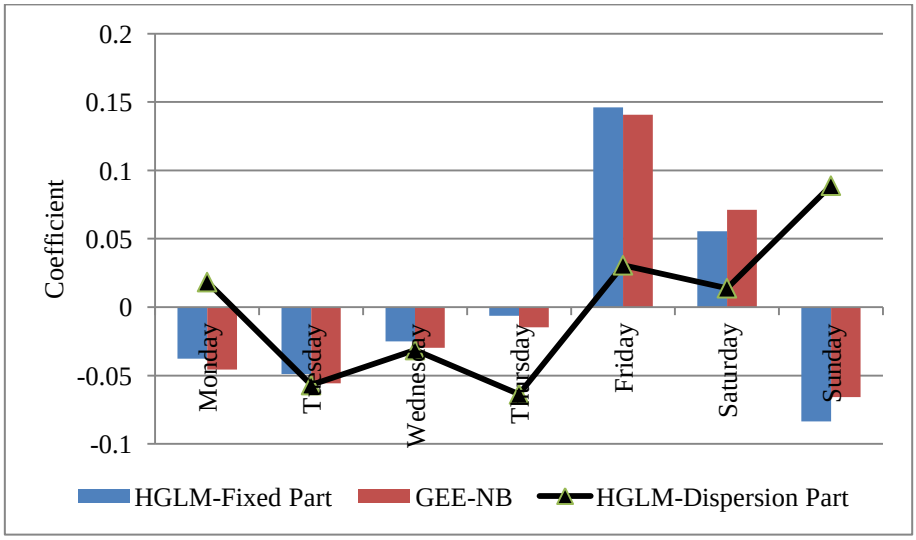
One of the main advantages of using regression analysis for the dispersion model is to test the significance of individual levels. The symbol ϕ represents the variation within each class of observations. In this model the age group, gender, month, and day of week variables were used in the dispersion part. The coefficients of dispersion models provide additional information by quantifying the amount of variation within the corresponding group.

The coefficients obtained from dispersion part of the HGLM are also shown in Figure 5.9. It is found that in the case of age groups, greater mean values tended to have greater dispersion

except for the age group under 17 where the mean is reduced and dispersion is highest. The under 17 age group had the lowest coefficient in the fixed part and had a higher coefficient in dispersion part which shows a lower number of casualties but it involves more variation in number of casualties. The coefficient of gender was found to be non-significant which is also evident from the Table 5.5; when Gender is removed from dispersion part, it decreased the EQL by only 3. In the case of day of the week, weekends have greater dispersion than weekdays with the exception of Friday and Monday. Monday had a reduced mean but a higher dispersion. Sunday had the lowest mean with the highest dispersion. In the case of month, there was no general relationship between variation in mean and dispersion. October, November and December had elevated mean values out of which November had reduced dispersion. All other months except these had reduced values of both mean and dispersion with the exception of January and June. November had a substantially elevated mean with reduced dispersion and December had substantially elevated values of both mean and dispersion.

Figure 5.9: Comparison of coefficients by HGLM and GEE-AR1 (Dataset 5: Car)





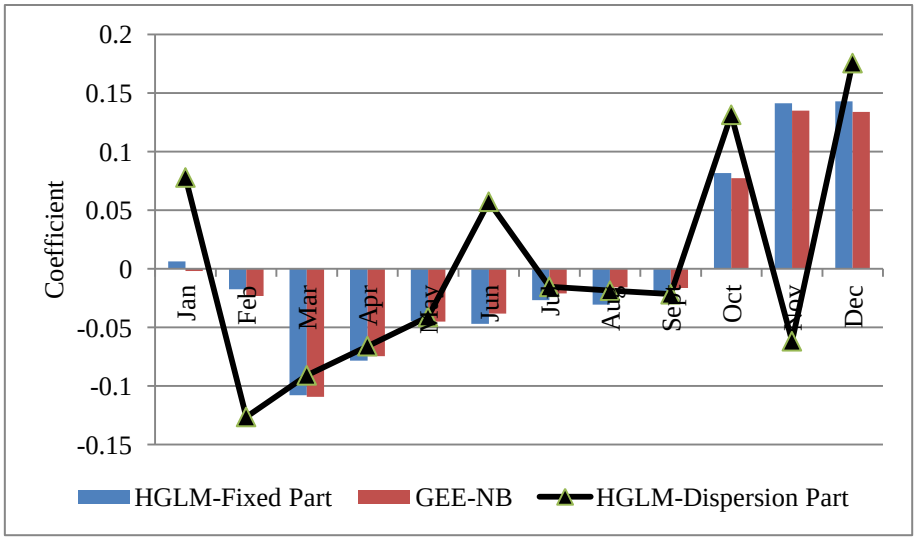


Table 5.8: Comparison of the coefficients and *t* values of the some variables by HGLM and GEE-AR1 (Dataset 5: Car)

Variable (Car)	**Coefficient HGLM**	***t* value HGLM**	**Coefficient GEE**	***t* value GEE**
Under 17	-0.913	-100.18	-0.907	-105.50
17-20	1.003	151.92	1.008	130.41
30-39	0.420	79.70	0.414	57.95
40-49	0.155	27.59	0.152	20.22
50-59	-0.123	-20.81	-0.126	-15.71
60-69	-0.467	-66.39	-0.468	-50.18
70 plus	-0.834	-115.37	-0.833	-89.73
Gender	0.023	13.11	0.024	10.79
Under 17.Male	-0.280	-20.89	-0.281	-22.60
17-20. Male	0.214	24.46	0.212	19.87
30-39. Male	-0.016	-2.17	-0.014	-1.35
40-49. Male	-0.075	-9.43	-0.073	-6.85
50-59. Male	-0.131	-15.37	-0.132	-11.51
60-69. Male	-0.013	-1.34	-0.015	-1.10
70 plus. Male	0.240	23.22	0.239	18.00
Monday	-0.038	-9.15	-0.046	-11.40
Tuesday	-0.049	-12.50	-0.056	-14.21
Wednesday	-0.025	-6.34	-0.030	-7.61
Thursday	*-0.006*	*-1.66*	*-0.015*	*-3.80*
Saturday	0.056	14.34	0.071	18.48
Sunday	-0.084	-19.69	-0.066	-16.68
January	*0.006*	*0.40*	*-0.002*	*-0.26*
February	*-0.018*	*-1.12*	-0.023	-3.23
March	-0.108	-6.87	-0.109	-15.59
April	-0.079	-4.99	-0.075	-10.54
May	-0.045	-2.84	-0.045	-6.50
June	-0.047	-2.97	-0.038	-5.47
July	*-0.027*	*-1.70*	-0.021	-3.07
August	*-0.031*	*-1.95*	-0.016	-2.29
September	*-0.019*	*-1.21*	-0.016	-2.35
October	0.082	5.18	0.077	11.40
December	0.143	9.02	0.134	19.43
Time	-9.6E-05	-10.57	-9.4E-05	-23.07
Holidays	-0.057	-9.19	-0.039	-6.44
New Year	-0.142	-7.61	-0.096	-5.94
Christmas	-0.196	-11.85	-0.180	-12.07
Constant	12.04	-464.97	-11.96	-556.97

Italics indicate the non-significant t values at the 5 percent level

Table 5.9: Significant coefficients of random part (Dataset 5: Car)

Month and Year	**Coefficient of month.year**	***t*-value of month.year**
Jan-01	0.045	2.34
Feb-04	-0.069	-3.67
Mar-03	-0.048	-2.55
Mar-04	0.040	2.17
May-01	-0.043	-2.21
July-03	-0.044	-2.32
Aug-03	-0.050	-2.67
Aug-04	0.047	2.49
Sept-02	-0.053	-2.78
Oct-02	0.059	3.19
Nov-02	0.040	2.31
Nov-04	-0.057	-3.04
Dec-03	-0.043	-2.28

5.6.1.5 Comparison of the estimated number of casualties by the HGLM and GEE-AR1 models

The number of casualties on each day for each group (age and gender combination) were estimated using each of the full model with HGLM Poisson-gamma distribution with a log link, and the GEE model with negative binomial regression and AR1 errors as shown in Section 5.5.2. These estimates were compared with the casualties observed on the corresponding days. It was observed that no particular difference was found between the number of casualties estimated by the HGLM and GEE models. The root mean square error (RMSE) for observed casualties and estimated by HGLM is 8.96 while for casualties estimated with GEE the value of RMSE is 9.05. The lowest value of RMSE is preferred.

It was generally observed from Figure 5.10 that the casualties estimated by both models fitted the observed data well as the line of equality passes through the centre of the data. However, there were some outliers which were not properly estimated by both the models. A few

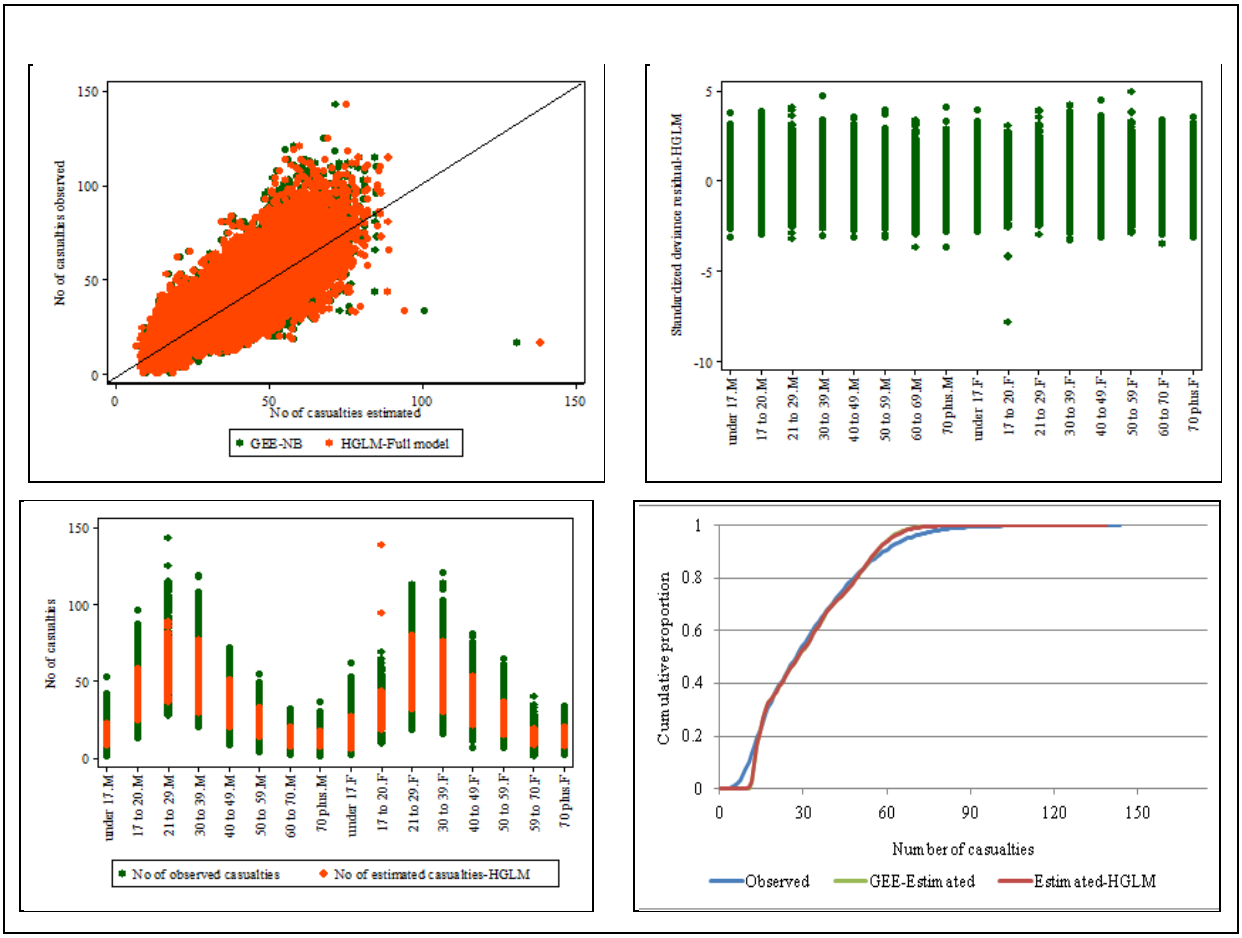
outliers were highlighted. It was observed that the highest number of 143 car casualties occurred on Saturday 12th October 2002 which belonged to the male 21 to 29 age group. The estimated casualties for this category were 75 by using HGLM and 71 by using GEE. Another outlier was for males of same age group on Friday 25th July 2003 with a value of 125 whereas the casualties estimated by the HGLM model were 69 and GEE predicted 67 car casualties. In the same way some observations were estimated with a higher value than the observed casualties. Upon detailed analysis of the data it was found that on 2nd January 2001 (Tuesday) only 17 casualties were observed for females of the age group 17 to 20 but HGLM and GEE estimated it to be 138 and 131 respectively. For the same age group, on 1st January 2001 (Monday) 34 casualties were observed but HGLM and GEE models estimated it to be 93 and 101 respectively.

From a standardized deviance residual graph, a few outliers with highest positive standardised deviance residual were identified which helped to identify those observations on which the number of casualties was estimated to be higher than the observed casualties. The graph shown below highlights that Monday 30th June 2003 (age group 50 to 59: females) and Thursday 1st February (age group 30 to 39: males) had the highest positive residuals of 4.93 and 4.75 respectively. The number of casualties estimated by the model on these two days was lower than the estimated casualties. Upon detailed investigation it was found that 65 female casualties in the 50 to 59 age group were observed on 30 June 2003 which was estimated to be only 23 casualties. Similarly 119 male casualties of age 31-39 years were observed on 1st February 2004 which was estimated to be 58 casualties by the HGLM model. This shows that HGLM model was not able to estimate high number of casualties that were observed during various periods of year. The most negative SDRs were found for 2nd and 1st January 2001. It was also observed from the graph of standardised deviance residuals that most of the SDRs lay between -4 and +4.

The third graph shows that there was less difference in the numbers of observed casualties between male and female. This was also evident from the results of the EQL; when gender was removed from the model it produced a reduction of only 3 which suggest that male and female are equally variable. Above the age of 30 the number of casualties decreases equally for male and female.

From the cumulative proportion graph it was observed that only 261 (2 percent) of observations had a value less than or equal to 6 casualties per day. However, the estimated model had only 1 observation in the same group. It can be seen that at values of less than or equal to 14 casualties per day both models had the same proportion of observations which is about 22 percent of the whole data. This suggests that observations lying on the tail were not estimated precisely by either model. Apart from this small change the observed cumulative proportion graph matched the estimated cumulative proportion graph.

Figure 5.10: Comparison of casualties observed and estimated, standardised deviance residuals produced by HGLM and GEE-AR1 (Dataset 5: Car)

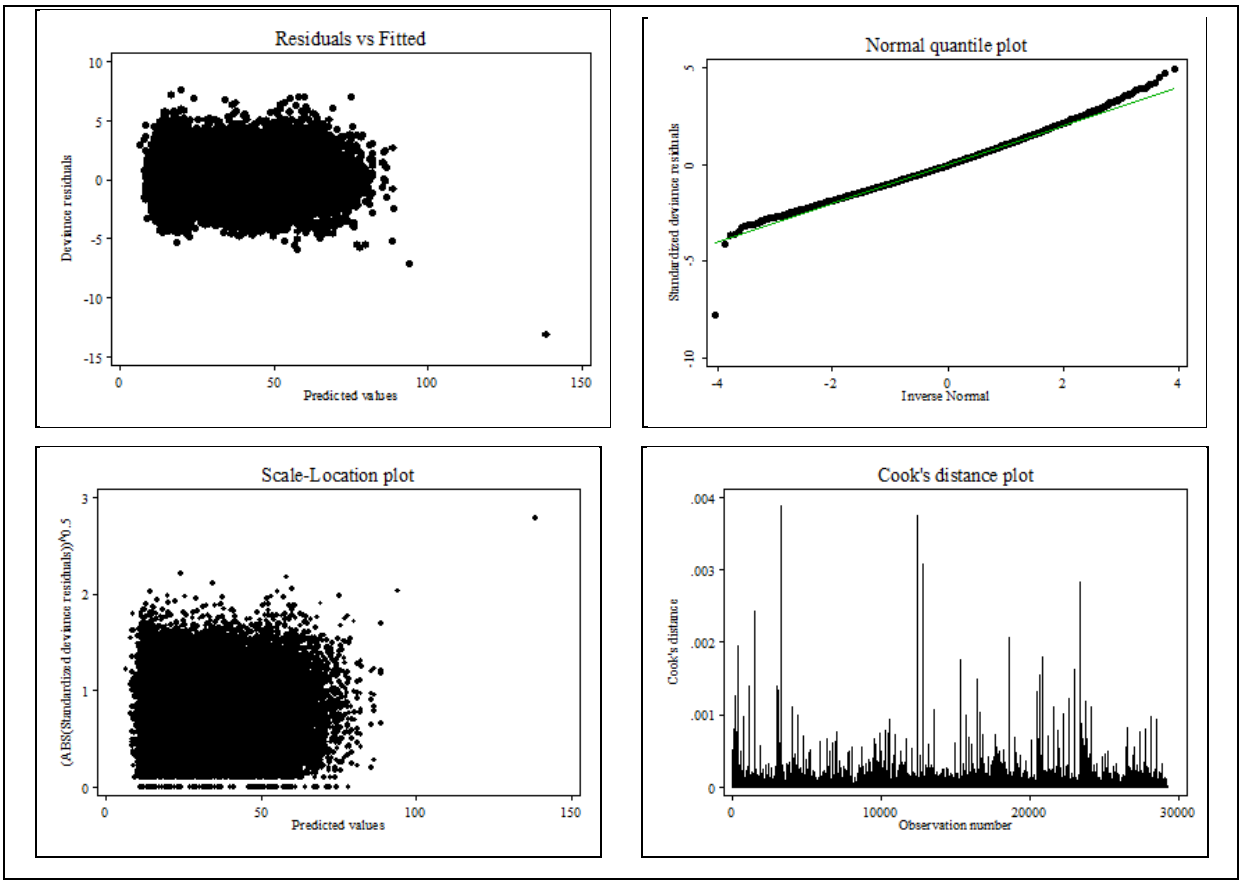


5.6.1.6 Final model checking graphs

In order to investigate the extent to which the trends in the data are represented in the HGLM model, the deviance residual were analysed. The graph of the deviance residuals against fitted values in Figure 5.11 show that deviance is scattered around the zero line. The normal quantile plot appears to be close to the line of equality which supports the assumption of

normality of residuals. However, beyond 3 the cumulative residual curve deviates slightly from the straight line suggesting that a few points are outliers. Scale-Location plot also shows that the deviance doesn't increase with mean. Cook's diagram shows that there were some strongly influential observations in the data: the results were investigated further to identify to which group these observations belonged. The 100 observations with highest Cook's distance were investigated. It is found that out of these 100 observations, 40 belonged to December, of which 18 were 25th December while the remaining 22 were 26th December. Most of observations with the highest Cook's distance belonged to 25th, 26th December which is always Christmas holidays. A further 24 observations belonged to January out of which 19 were 1st January which was a New-Year holiday. This shows that these days of the year vary from year to year in a way that is not captured fully by the present model. As the Cook's distance value for each of the observation was less than 1, so all of these observations were kept in the original datasets to reflect the actual variation in number of casualties on these days. Heteroscedasticity tests were not undertaken here because in these HGLM models, unequal variances were accommodated by using the DTERMS.

Figure 5.11: Diagnostic plots: Full model-HGLM (Dataset 5: Car)



5.6.1.7 Estimating the car casualty rate per million population:

The age and gender specific rate of casualty per million population was calculated by dividing the estimated number of casualties occurring on each day by population of that age and gender group. The ratios presented here represent the annual age and gender specific rate of casualty per million person years. It is to note that these rates do not take account of the exposure in terms of distance travelled. As noted by Legge et al (1998) if the distance travelled were taken into account instead of population, the estimated age and gender profiles would be different from those presented in Table 5.10.

The numbers of car casualties shown in Figure 5.10 were estimated by the HGLM model. It is observed that Females had higher car casualties in under 17 age group and at ages greater than 40 years than males of same age group. The number of car casualties were found to decrease with increase in age after 30 years of age. Table 5.10 which also shows the rates of car casualty per million person-years further reveals that;

- Females had a higher rate per million person-years of being a car casualty than males in all age groups except in the 17 to 39 age group and the 60 plus age group.
- Persons in the 17 to 20 age group had the highest rate per person-year of being a car casualty among all age groups.
- Above the 17 to 20 age group the car casualty rate per person-year decreases with increase in age. The Under 17 group have lowest car casualty rate per person-years among all age groups.

Table 5.10: Number of car casualties estimated by HGLM and estimated car casualty rate per million person-years

Estimated number of casualties by the model								
Age group	Under 17	17-20	21-29	30-39	40-49	50-59	60-69	70+
Male	15.4	41.5	60.8	53.0	35.4	23.3	13.0	12.7
	(15)	*(42)*	*(61)*	*(53)*	*(35)*	*(23)*	*(13)*	*(13)*
Female	18.4	30.4	54.4	52.3	37.1	25.9	13.6	13.8
	(19)	*(31)*	*(54)*	*(52)*	*(37)*	*(26)*	*(14)*	*(14)*
Rate per million person-years								
Male	24.9	276.6	186.1	122.7	88.7	63.5	50.6	45.2
Female	31.4	213.0	167.3	119.0	91.3	69.14	49.0	34.0

( )represents the observed number of casualties from the data

5.6.2. Model selection process, goodness of fit and model checks for Datasets 6-9 (Walk, Bicycle, Motorcycle and Bus)

This section shows the results of the models developed, goodness of fit of the preferred model and various checks to validate the model for each of the walk, bicycle, motorcycle and bus casualty datasets (Datasets 6 to 9). The same procedure of model development as described in section 5.5.2 was used for each of these. In this section the results of the each dataset are presented together with the aim of achieving any common and distinct features among these datasets.

5.6.2.1 Hierarchical generalized linear model (HGLM) Walk-Bicycle-Motorcycle and Bus Datasets 6 to 9:

The Hierarchical Generalized Linear Model (HGLM) with Poisson-gamma distribution with log link was used for data on each of the walk, bicycle, motorcycle and bus casualties (Datasets 6 to 9). In the first step, the full model as shown in section 5.5.2 was developed. After this, individual variables were removed from the model to investigate their partial effect on the model fit as reflected in the h-likelihood values.

From Table 5.11 it was found that:

- Day of the week had the highest effect among all other variables in the fixed part in each of the casualty data, although it was found to be less sensitive in the motorcycle data.
- The age.gender variable was found to be more sensitive in the car casualty data in comparison to all other casualty datasets.
- Month had the uniform effect on the model fit in all of the casualty data except for bus where it had the least effect.
- The time variable had the least effect on the bicycle casualty data.
- Public holiday had the least effect on motorcycle data while new-year holidays had the greatest effect on walk casualty data.
- Christmas holidays had the highest effect on motorcycle data.

In the second step, the month.year interaction was removed from the random part of the full model, keeping the same fixed and dispersion parts. It is observed that removal of month.year reduced the h-likelihood of the fixed part confirming that it contributes substantially to model performance. The highest effect of 375 was observed in walk (Dataset 6) while the lowest effect of 6 in the h-likelihood was observed in bus (Dataset 9).

In the third step, the variables were removed in turn from the dispersion part of the model while keeping the same fixed and random parts. The results showed that age group is the most important variable which affected the extended quasi likelihood (EQL) of the models. The highest reduction of 1,226 in EQL was observed for motorcycle (Dataset 8) while lowest effect of 478 was observed for bicycle (Dataset 7). The removal of Gender had the least effect of 3 on the EQL with 1 degree of freedom in the car (Dataset 5). This suggests that Gender is equally variable in this dataset. In the same way, day of the week had comparatively less effect of only 3 on EQL for bicycle data. This suggests that there is no additional structure in the variability in this case which can be represented through day of week.

The coefficient λ whose exponential value represents the variance of the random effects, in the present case, month and year is found to be significantly different from 0 among all modes. The value of λ ranged from -5 to -7. The h-likelihood values obtained for the fixed, random and dispersion parts are shown in Table 5.11.

5.6.2.2 Analysing the temporal effects

In this section the temporal effects were analysed for the dataset 6-9 (Walk, Bicycle, Motorcycle and Bus). For this, each of the full models as shown in section 5.6.2.1 was used. The same procedure as used in section 5.6.1.2 was used.

From the results shown in Appendix Table A5.4 it was observed that in each case (each mode) the addition of square of time variable to the fixed part of the model has not resulted in substantial improvement in the h-likelihood. The h-likelihood improved by value of only 1 when square of time was included to the full model for Walk and Bus data. The estimated *t* value of square of time had non-significant *t* values whilst it had high VIF (in range of 16) showing multicollinearity with other variables present in the model.

Similarly, in the dataset of Bicycle and Motorcycle the addition of square of time improved the h-likelihood of the model only by value of 7 and 4 respectively. The *t* values of the square of time were found to be significant with value of 3.51 and -2. 38 for Bicycle and Motorcycle data respectively. The VIF of the square of time was 16 in each case which was higher than the acceptable value which showed the presence of multicollinearity.

Slight improvement in h-likelihood, non-significant *t* values of time square, and high VIF values shows that there is no substantial temporal effect remaining in each of these models that can be represented by the quadratic terms.

Table 5.11: Results of h-likelihood (Walk, Bicycle, Motorcycle and Bus: Datasets 6 to 9)

Models	**Variables**	**d.f.**	**H likelihood and change in its value**				
			Car	**Walk**	**Bicycle**	**MC**	**Bus**
FIXED	**Full Model**		**201,681**	**134,387**	**97,079**	**102,595**	**82,559**
	- Age. Gender	7	+1,664	+314	+318	+126	+437
	- Day of week	6	+1,840	+1,823	+2,395	+281	+1,310
	- Month	11	+111	+144	+107	+133	+48
	- Time	1	+72	+95	+23	+49	+44
	- Holiday	1	+86	+121	+180	+5	+171
	- New-Year	1	+151	+310	+56	+111	+12
	- Christmas	1	+62	+96	+56	+207	+51
RANDOM	**Full Model**		**201,257**	**134,021**	**96,796**	**102,264**	**82,280**
	- Month.Year	1	+234	+375	+162	+135	+6
DTERMS	**Full Model**		**201,708**	**134,489**	**98,244**	**101,264**	**85,697**
	- Age	7	+1,164	+1,183	+478	+1,226	+855
	- Gender	1	+3	+49	+166	+613	+42
	- Month	11	+113	+100	+54	+126	+24
	- *Day of week*	6	+33	+93	+3	+40	+119
Coefficient	λ		-6.73	-6.98	-5.17	-5.48	-5.3
			(0.001)	*(0.0009)*	*(0.005)*	*(0.004)*	*(0.004)*
t value			-30.7	-26.6	-22.9	-24.4	-21.4

( ) italic represents the variance of the random part of the model

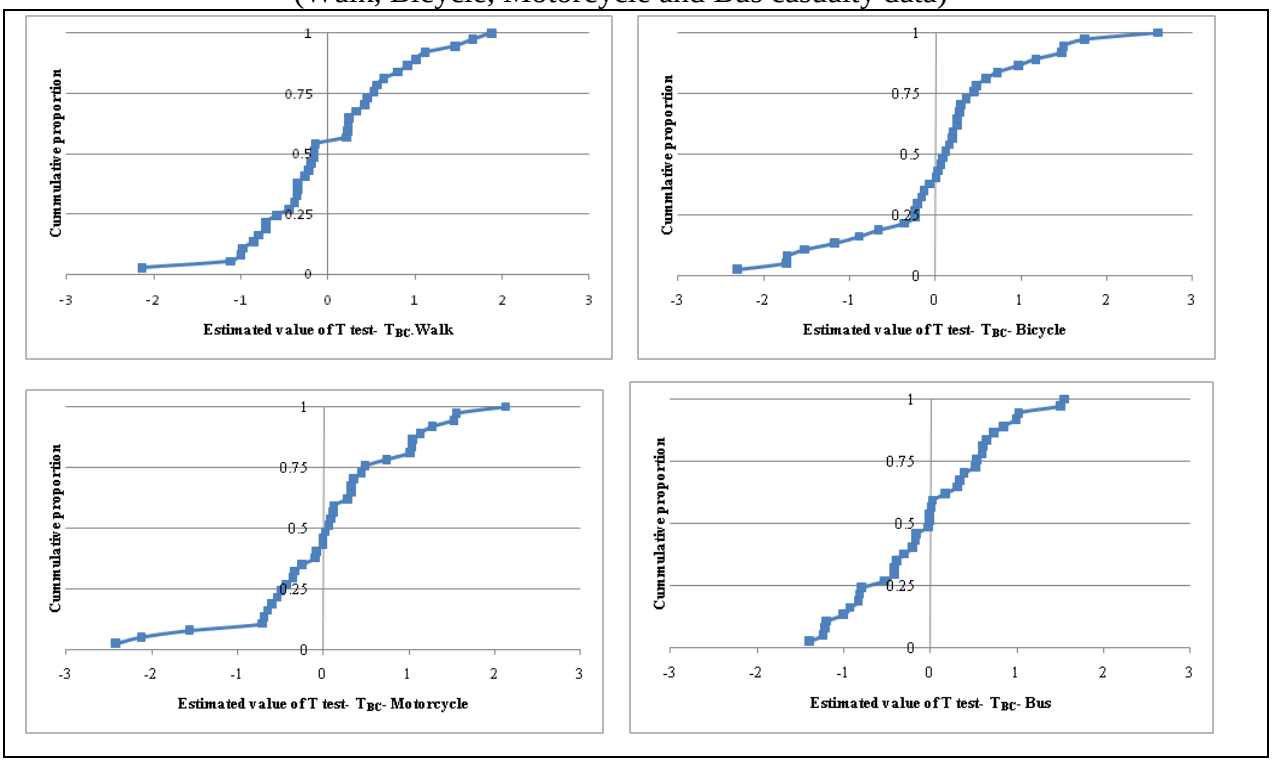
5.6.2.3 Split sample tests

In order to check the consistency of the model parameters, split sample validation tests were carried out. To do this, the dataset was randomly partitioned into two, each part with 14,608 observations. GenStat software was used to estimate the model parameters of Datasets B and C which were then compared. After this the coefficients of the fixed part were compared between models B and C for each of the casualty datasets (Datasets 6 to 9). The *T* test was used to compare the coefficients of Datasets B and C. T_{BC} values were estimated using the formula 2-32. It is found from *T* test values that the coefficients of model B are not significantly different from the coefficients of model C with the exception of few variables as most of the estimated values of T_{BC} are less than 1.96. The comparison of coefficients, *t* values and estimated *T* test values T_{BC} for each dataset are shown in Appendix Tables A5.5 to A5.8 and Appendix Figures A5.1 to A5.4. From the estimated values of T_{BC} it was found that the coefficients of the following variables had changed, it is to note that each model has 37 coefficients:

- In the walk casualty data, only the coefficient for Thursday had changed.
- In the bicycle casualty data only the coefficients of gender and public holidays had changed.
- In the motorcycle data the coefficient of age group 70 plus, 70 plus.male and Wednesday had changed.

The estimated values of the T_{BC} with the cumulative proportion for each of the casualty data modes are shown in Figure 5.12 which clearly shows that there were very few variables for which the magnitude of the coefficients had changed (Models B and C) in each of the casualty data modes while in most of the cases the estimated value of the T_{BC} was between -1.96 and +1.96 which is an indication that the value of the coefficient has not changed.

Figure 5.12: Estimated values of T_{BC} with cumulative proportion in Dataset 6 to 9 (Walk, Bicycle, Motorcycle and Bus casualty data)



5.6.2.4 Comparison of Coefficients (HGLM and GEE-AR1; Car-Walk-Bicycle-Motorcycle-Bus)

a. Fixed Part:

Due to the reasons explained in section 5.6.1.4, we will only look into the similarities in the estimated coefficients (similar signs and values of estimated coefficients) and differences in the estimated coefficients. The coefficients of HGLM with Poisson-gamma distribution and log link, and GEE-AR1 with negative binomial were compared for each of the walk, bicycle, motorcycle and bus casualty data. It was found that the coefficients that were significant at the 95 percent level in HGLM and GEE-AR1 models had the same sign, though a slight change was observed in the *t* values of the variables. The individual results of each mode are shown in the appendix Tables A5.9 to A5.12 and appendix Figures A5.5 to A5.8.

In this section the coefficients used in the fixed part in each of the casualty data (car, walk, bicycle, motorcycle and bus) are compared to identify any similar patterns in the data. It is observed from Figure 5.13 that:

- The 17 to 20 age group had greatest casualty rate per person-years for car and motorcycle casualties. Among bicyclist casualties, the 21 to 29 age group had the greatest rate. The Under 17 age group had the highest casualty rate per person-years in walking while older people (70 plus) have greatest rate per person-years while travelling in bus.

- For all modes except walk and bus the casualty rate per person-years of getting injured in road accident decreases after a certain age group. For walk, the casualty rate decreases with increase in age but after 50 years it increases again. The rate per person-years for getting injured while travelling in bus increases after the age of 40 years.

- Among car casualties, those under 17 age had least casualty rate per person-years for getting injured in road accident. Among bicyclist and motorcyclist, older people have the least casualty rate. The age groups 50 to 59 and 30 to 39 had less rate of getting injured while walking and travelling in bus respectively.

- Friday is associated with the greatest casualty rate than any other day of the week in each of the modes except bicyclists where Wednesday was found to have the highest rate. Weekdays have greater casualty rate than weekends but car travellers have greater rate on Saturday than weekdays except Friday.

- In summer months (June, July, August and September) bicycle and motorcycle had greatest casualty rate whereas in winter months (December, January and February) they have lowest rate.

- Car and Walk modes have greater casualty rate per person in November and December while car has lowest casualty rate in March and walk has lowest in August.

b. Random Part:

The interaction of month.year was used in the random part of the HGLM model. The exponential of the parameter λ represents the variation between corresponding months in different years. It was found that λ had the coefficient ranging from -5 to -7 with a

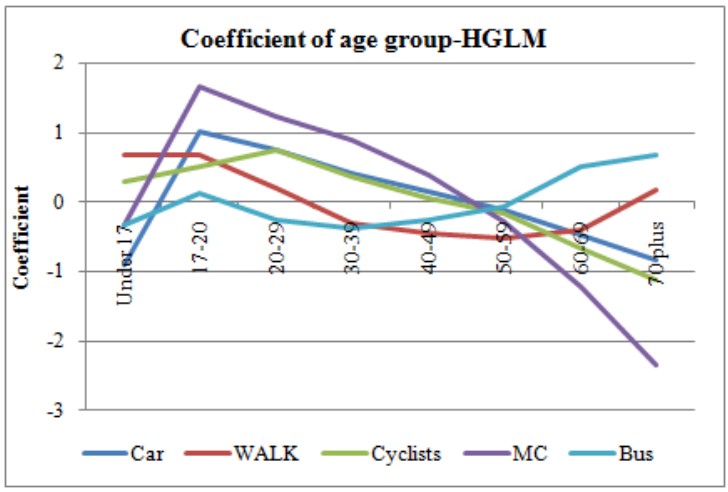
significant *t* value for each mode. Only 42 out of the 300 observations (each mode had 60 coefficients) of month.year interactions had significant *t* values which show that these months in the mentioned years were significantly different from zero. The interpretation of the value of the random term can be made that January 2002 was different and had fewer walking casualties than was usual in January. In the same way August 2002 was different but it had higher motorcycle casualties than usual in August as it has a positive coefficient. The detailed results of significant coefficients of the random part for each mode are shown in Table 5.12.

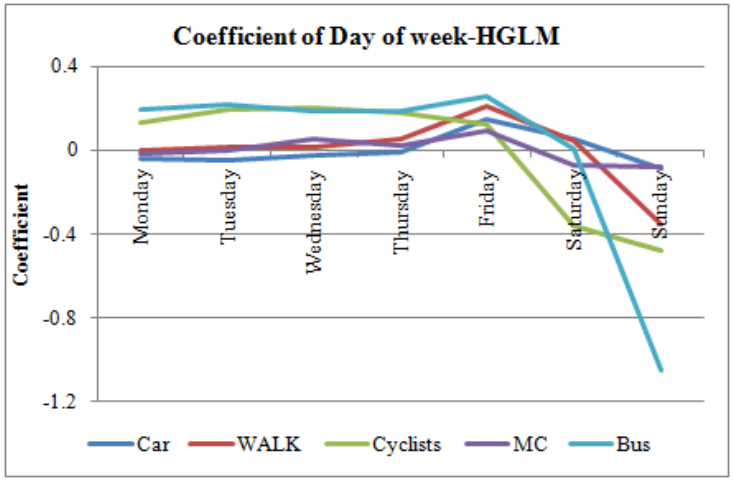
c. Dispersion Part:

One of the main advantages of using regression analysis for the dispersion model is to test the significance of individual levels. The symbol ϕ represents variation within each class of observations. In this model the age group, gender, month, and day of week variables were used in the dispersion part. The coefficients of dispersion models provide additional information by quantifying the amount of variation within the corresponding group. The individual results for each mode are shown in Appendix Figures A5.5 to A5.8. In this section, the coefficients used in the dispersion part of the each casualty data (car, walk, bicycle, motorcycle and bus) from the full model are compared and are shown in Figure 5.14. The results are summarised below:

- The under 17 age group had a greater variation in all modes except motorcycle where 30 to 39 years had the highest variation.
- For car the dispersion decreases with increase in age after the 17 to 20 years age group whereas for motorcyclists it decreases after 30-39 years of age.
- For walk and bus modes the elderly group (70+) had a greater variation than all other age groups except those under 17.
- Sunday had the highest variation among all days of the week for all modes except bus where it had the least dispersion. Bus casualties had greatest variation on Tuesday.
- December had the greatest variation among all months for each of the casualty data except motorcycle and bus which had greater variation in June.

Figure 5.13: Comparison of coefficients from Fixed part of Model (Datasets 5-9)





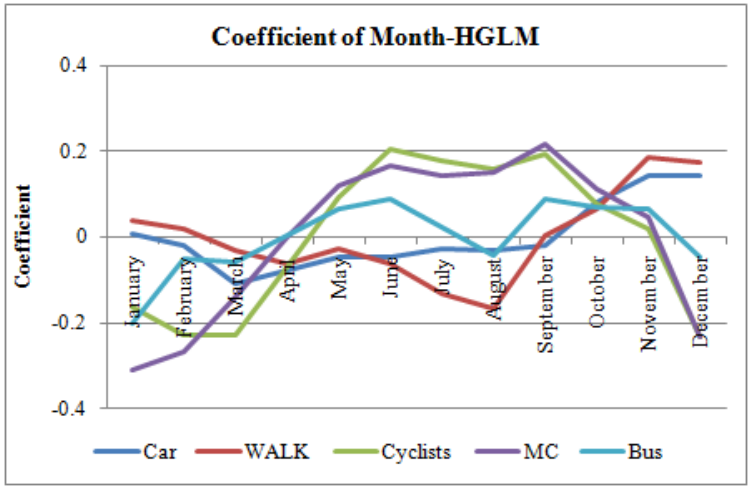
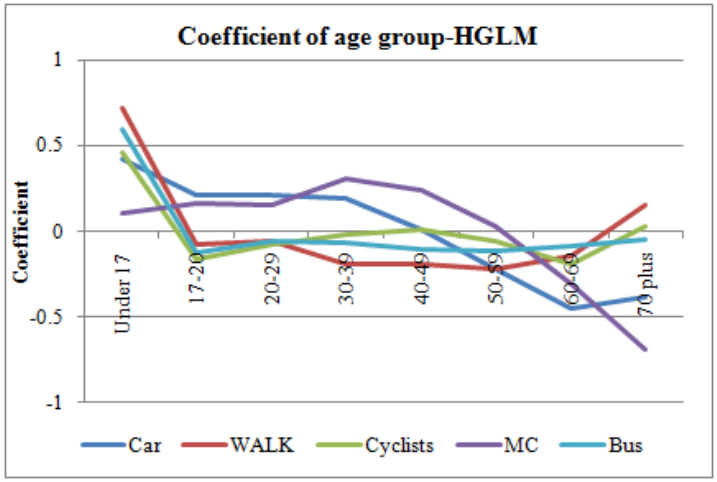


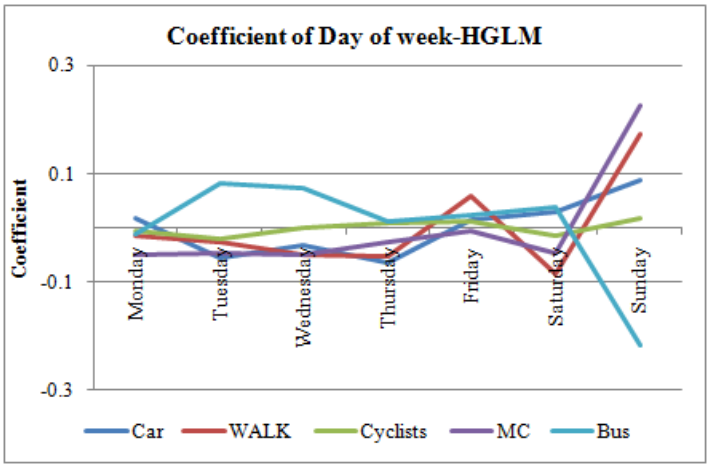
Table 5.12: Comparison of significant coefficients from Random part of Model (Datasets 5-9)

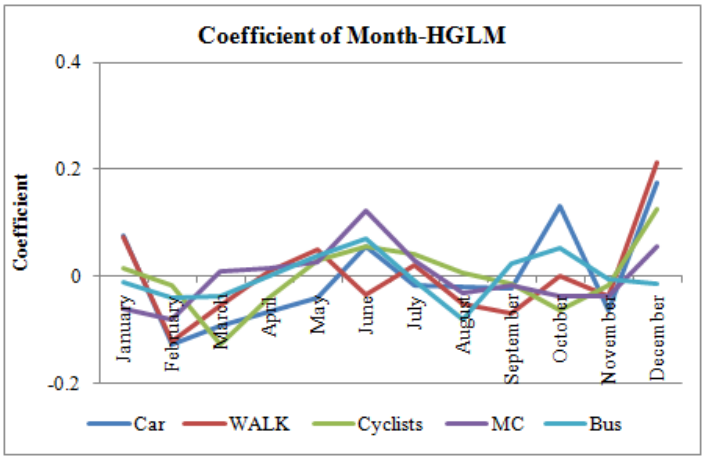
	Car		Walk		Bicycle		Motorcycle		Bus	
Variable	Coefficient	*t* value	Coefficient	*t* value	Coefficient	*t* value	Coefficient	*t* value	Coefficient	*t* value
Jan-01	0.045	2.34	-	-	0.137	3.1	-	-	-	-
Jan-02	-	-	-0.046	-2.09	-	-	-	-	-0.098	-1.96
Jan-03	-	-	-	-	-0.099	-2.1	-	-	-	-
Feb-04	-0.069	-3.67	-	-	-	-	-	-	0.100	2.17
Mar-01	-	-	-	-	-	-	-0.116	-2.93	-	-
Mar-03	-0.048	-2.55	-	-	0.089	2.06	0.158	4.36	-	-
Mar-04	0.04	2.17	-	-	-	-	-	-	-	-
Apr-01	-	-	-	-	-	-	-0.117	-3	-	-
Apr-02	-	-	-	-	-	-	0.092	2.53	-	-
May-01	-0.043	-2.21	-	-	0.147	3.44	-	-	-	-
May-02	-	-	-	-	-0.091	-2.09	-	-	-0.099	-2.09
May-03	-	-	-0.055	-2.48	0.162	-3.67	-	-	-	-
May-04	-	-	-	-	0.091	2.14	-	-	-	-
Jun-02	-	-	-	-	-0.14	-3.19	-	-	-	-
Jul-03	-0.044	-2.32	-	-	-	-	-	-	-	-
Aug-02	-	-	-	-	-0.096	-2.22	0.102	2.94	-	-
Aug-03	-0.05	-2.67	-	-	-	-	-	-	-	-
Aug-04	0.047	2.49	-	-	-	-	-	-	-	-
Sep-01	-	-	-	-	-0.103	-2.32	0.085	-2.25	-	-
Sep-02	-0.053	-2.78	-	-	-	-	-	-	0.164	3.75
Oct-02	0.059	3.19	-	-	-	-	-	-	-	-
Oct-04	-	-	0.048	2.23	-	-	-	-	-	-
Nov-02	0.04	2.31	-	-	-	-	-	-	-	-
Nov-04	-0.057	-3.04	-	-	-	-	-	-	-	-
Dec-03	-0.043	-2.28	-	-	-	-	-	-	-	-

It is to note that only significant random variables are shown in table. Full details of models are shown in Appendix table A5.11 to A5.15

Figure 5.14: Comparison of coefficients from dispersion part of Model (Datasets 5-9)







5.6.2.4 Comparison of the estimated number of casualties by HGLM and GEE-AR1 model for (Walk, Bicycle, Motorcycle and Bus casualty data) Datasets 6 to 9

The number of casualties on each day for each group (age and gender combination) were estimated using each of the full model with the HGLM Poisson-gamma distribution with a log link, and the GEE model with negative binomial regression and AR1 errors as shown in Section 5.5.2. These estimates were compared with the casualties observed on the corresponding days in each of the casualty data modes. The root mean square error (RMSE) was estimated for each of the casualty data which showed in Table 5.13 that there was very little difference between the observed casualty numbers and those estimated by HGLM and GEE.

It is observed from the graphs in Figures A5.9 to A5.12 shown in appendix that no particular difference was found between the number of casualties estimated by HGLM and GEE models. Both models (HGLM and GEE) fitted the observed data well in each casualty data (Datasets 6 to 9) as the line of equality passed through the centre of the data. However there were some outliers which were not properly estimated by both the models. It was also observed from the graphs that most of the standardised deviance residuals lay between -4 and +4. The cumulative proportion graph for the estimated number of casualties for HGLM and GEE were mostly identical whereas the observations lying on the tail were not estimated precisely by both models. Except for this small change, the estimated cumulative proportion graph matched the observed cumulative proportion graph in each casualty data mode.

Table 5.13: Root mean square values of the casualty data (Walk, Bicycle, Motorcycle, Bus-Dataset 6-9)

	Root mean Square	
Casualty data	Observed and HGLM	Observed and GEE
Walk	3.32	3.33
Bicycle	2.06	2.09
Motorcycle	2.73	2.76
Bus	1.65	1.65

5.6.2.6 Final Model checking graphs

The deviance residuals were evaluated to investigate the extent to which the trends in each of the sets of casualty data (Datasets 6 to 9) are represented in the HGLM model. The graphs are shown in Appendix Figures A5.13 to A5.16. The deviance residuals against fitted values and scale location plot shows that HGLM in each of these dataset had not absorbed the entire trend. The graph does show that deviance is scattered around the zero line. The normal quantile plot appears to be on the straight line which supports the assumption of normality of residuals. However, in each case the residuals slightly deviated from a straight line near the tails suggesting a few points are outliers. Cook's diagram showed that there were influential observations in the data. Most of these observations belonged to December (25^{th} and 26^{th} December) and 1^{st} January which is a New-Year holiday. However, the Cook's distance value was less than 1 in each case. Due to this, all of these observations were retained in the original datasets to reflect the actual variation in number of casualties on these days. Heteroscedasticity tests were not undertaken here because in these HGLM models, unequal variances were accommodated by using the DTERMS.

5.6.2.7 Estimating the casualty rate per million population (Datasets 6 to 9)

The age and gender specific rate of casualty per million population was calculated by dividing the estimated number of casualties occurring on each day by the national population of that age and gender group. The number of casualties for each of the datasets was estimated by the HGLM model. This approach was followed in order to account for the relationships among the explanatory variables that are present in the dataset, which represents the structure of the observations. Table 5.14 shows the estimated casualty rate per million population, the number of casualties estimated by HGLM for each age group by gender, and the observed casualty data. These ratios represent the annual age and gender-specific casualty rate per million person-years in the national population as reflected in the HGLM model. It is to note that because these values do not take account of the exposure in terms of distance travelled, they can not be interpreted in terms of risk per vehicle-kilometres. From Table 5.14 it is found that:

- Car users had the highest casualty rate per person-years in comparison to walk, bicycle, motorcycle and bus users with the exception of males under 17 years of age, who had greatest casualty rate in walking.
- Persons in the 17 to 20 age group had the highest casualty rate per person-years among all age groups except in bus casualties where older people (70+) had the highest rate.
- The casualty rate per person-years decreased with increasing age in each of car, walk, bicycle and motorcycle except for elderly people (70+), at which age there is an increase in casualty rate for walking.
- Males had a higher casualty rate per person-years than females in walk, bicycle and motorcycle data.
- Females had a higher casualty rate per person-years than males for all age groups in bus data, and also in the under 17 and 40 to 60 age groups for car users.

5.7 CONCLUSION

The purpose of this part of the study was to investigate the use of the Hierarchical Generalized Linear Model (HGLM) having Poisson-gamma distribution with logarithmic link for the analysis of road accident casualty data. As part of this, comparison was made with a generalized estimation equation (GEE) model with negative binomial distribution and AR1 time-series error structure. A further objective was to explore the possibilities for the combined use of accident and casualty information from the national accident data in conjunction with population data. It was found that casualty data in this case had some structure that could be identified by the HGLM. In order to draw inferences from this, models with random effects should be used which can incorporate the heterogeneities among the observations. In this application the HGLM model was preferred over the GEE-AR1 model due to the additional capability it offered in incorporating random effects and modelling the mean and dispersion jointly. Use of these capabilities is justified by the substantial improvement in model fit that is achieved. However, unlike the GEE-AR1, HGLM cannot accommodate time series error structure, which has affected estimates of coefficients of some of the variables. Magnitude and *t* values of the estimated coefficients differ between HGLM and GEE-AR1. The coefficients of age group, gender, interaction of age group and gender were similar in both models whereas the coefficients of day of week, month, public holidays,

new-year and Christmas holidays differed between the models. Generally in HGLM model, age group, gender, and interaction of age group and gender had greater *t* values whilst day of week and month had greater *t* values in GEE-AR1 model.

The age and gender-specific estimates of rate per million population-years presented here show how the casualty rate varies through the national population. These ratios take account of population rather than distance travelled. They therefore represent the rate of casualty for each age and gender group in the population. The estimated casualty rates would vary if different exposure variables were considered.

From the modelling results it was found that on a day the chance of casualty for a member of the population in each mode of travel varies according to the combination of their age and gender, the day of the week, the calendar month, and whether or not the day was a holiday or part of Christmas or New-Year holidays. The casualty rate on comparable days decreased by about 4 percent for each year during the period 2001-2005 for which data were analysed. The calendar month effect was found to vary at random from year to year. The dispersion of observations around their modelled values was found to vary systematically according to age, gender, calendar month and day of the week; this influenced the accuracy of estimates accordingly. From this dispersion modelling it was found that Sunday had more variation in the number of casualties than other days of the week. The under 17 age group was found to be more variable than other age groups. In the same way the oldest age group (70 plus) had more variation than other age groups in all modes except motorcycle where this group has least variation.

From the estimated results of number of casualties occurring on each day it was found that for each age group, males had higher or equal number of casualties than females in walk, bicycle, and motorcycle modes. For bus travel, females had more casualties than males in the same age group. The age group 17 to 20 years has the greatest number of casualties.

From these estimates it was found that the greatest casualty rate per million person-years among the modes arises in car use. People aged 17 to 20 years had greater rate per million person years of being a road casualty in all modes except bus where those aged 70 plus had the greatest rate. Males have a greater casualty rate than do females in all modes except bus, where females of all ages have a greater rate, and in car use either aged under 17 or between

40 and 59 where again casualty rate is greater for females. It is also found that the casualty rate per million person-years decreases with increase in age in all modes except walking where elderly people (70+) for whom the rate is greater than for those aged 30 to 59.

It is concluded from the estimates of casualty rate results per million population that this varies with age and gender for different modes. Suitable remedial policies, such as education and enforcement, could be addressed to target groups based on either higher number of casualties or on having higher casualty rate per person-year.

Table 5.14: Number of casualties estimated by HGLM and estimated casualty rate per million population

Age group	**Modes of travel**									
	Car		**Walk**		**Bicycle**		**MC**		**Bus**	
	M	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
<17	**24.9**	**31.4**	**31.6**	**23.4**	**17.71**	**3.5**	**10.6**	**1.3**	**2.5**	**3.3**
	(15.4)	(18.4)	(19.5)	(13.7)	(11.0)	(2.0)	(6.5)	(0.8)	(1.6)	(2.0)
	15	*19*	*19*	*14*	*11*	*2*	*6*	*1*	*2*	*2*
17-20	**276.6**	**213.0**	**32.9**	**23.4**	**20.0**	**4.4**	**80.6**	**9.7**	**2.6**	**5.2**
	(41.5)	(30.8)	(4.9)	(3.4)	(3.0)	(0.6)	(12.0)	(1.4)	(0.4)	(0.7)
	42	*31*	*5*	*3*	*3*	*1*	*12*	*1*	*0*	*1*
21-29	**186.1**	**167.3**	**22.3**	**14.5**	**16.9**	**5.6**	**41.9**	**6.3**	**2.6**	**3.5**
	(60.8)	(54.4)	(7.3)	(4.4)	(5.5)	(1.8)	(13.7)	(2.1)	(0.8)	(1.1)
	61	*54*	*7*	*5*	*6*	*2*	*14*	*2*	*1*	*1*
30-39	**122.7**	**119**	**15.5**	**8.7**	**15.8**	**3.7**	**39.4**	**4.6**	**2.9**	**3.2**
	(53.0)	(52.3)	(6.7)	(3.8)	(6.8)	(1.6)	(17.0)	(2.0)	(1.3)	(1.4)
	53	*52*	*(7)*	*4*	*7*	*2*	*17*	*2*	*1*	*1*
40-49	**88.7**	**91.3**	**11.7**	**7.7**	**11.4**	**2.7**	**25.2**	**2.7**	**2.7**	**3.5**
	(35.4)	(37.1)	(4.7)	(3.1)	(4.6)	(1.1)	(10.1)	(1.1)	(1.1)	(1.4)
	35	*37*	*5*	*3*	*5*	*1*	*10*	*1*	*1*	*1*
50-59	**63.5**	**69.1**	**9.2**	**7.0**	**7.3**	**2.2**	**11.5**	**1.4**	**2.5**	**4.3**
	(23.3)	(25.9)	(3.4)	(2.6)	(2.7)	(0.8)	(4.2)	(0.5)	(0.9)	(1.6)
	23	*26*	*3*	*3*	*3*	*1*	*4*	*1*	*1*	*2*
60-69	**50.6**	**49**	**9.6**	**8.0**	**4.5**	**1.3**	**4.3**	**0.5**	**2.9**	**7.0**
	(13.0)	(13.6)	(2.5)	(2.2)	(1.2)	(0.4)	(1.1)	(0.2)	(0.7)	(2.1)
	13	*14*	*2*	*2*	*1*	*0*	*1*	*0*	*1*	*2*
70+	**45.2**	**34**	**18.8**	**14.1**	**5.0**	**0.9**	**1.2**	**0.2**	**4.0**	**9.1**
	(12.7)	(13.8)	(5.3)	(6.0)	(1.4)	(0.3)	(0.3)	(0.1)	(1.1)	(3.7)
	13	*14*	*5*	*6*	*1*	*0*	*0*	*0*	*1*	*4*

Bold represents the casualty rate per million person-year
Values in brackets show the number of casualties estimated by HGLM
Italics show the number of observed casualties

6. SUMMARY AND CONCLUSIONS

This study has investigated the occurrence of road traffic accidents at a national scale. The four different national datasets of STATS 19, national travel survey data (NTS), population, and meteorological data for Great Britain over the 15 years (1991-2005) were analysed individually and jointly. Various statistical techniques were used including generalized linear model (GLM), generalized estimation equation (GEE), and hierarchical generalized (HGLM) linear model. The objectives of this investigation were to make joint use of these national datasets, determine the relationship between the number of road accidents and different variables available in the national datasets, evaluate the performance of GLM, GEE and HGLM models for this work, and to estimate and compare the risk of road accident involvement for various groups of road users on different kinds of roads. The following conclusions were drawn according to the set of objectives:

6.1 JOINT USE OF NATIONAL DATASETS

The national statistical datasets are valuable resources for road safety research, especially when used jointly. However, joint use of the four national datasets showed that these datasets are not immediately compatible with each other. The process of extracting the information for road accidents occurring on each day from the STATS 19 data (Accident section) was straightforward. The combined use of different sections of STATS 19 data is challenging as the Casualties and Vehicle sections on their own do not include information when and where the road accident occurred, whereas Accident section does not have information about road class, vehicle type, age and gender. The combined use of accidents, casualties and vehicle data from STATS 19 jointly with other sources such as traffic flow, meteorological and population data presents challenges to users because of difficulties in matching in temporal and spatial domains. Each of these national datasets had distinct road class, vehicle type and age groups which were reconciled with STATS 19 data for this study. A procedure was designed to combine the various national datasets including different sections of STATS 19 data for the modelling of road accidents, vehicles involved in road accidents and number of casualties which can be applied generally to national datasets of these kinds. Various datasets were developed in this study which linked the various sources of information and can be readily used for modelling.

6.2 RELATIONSHIP OF DIFFERENT VARIABLES TO NUMBER OF ROAD ACCIDENTS

The second objective was to identify the factors associated with variations in the risk of road accident occurrence per unit of distance travelled. Distance travelled on each day was used to represent exposure to risk: this was profiled by day of week and month of year by applying correction factors obtained from the Department for Transport to account for the day to day variation in distance travelled.

From the modelling results a clear difference was observed between the risk estimated for weekday and each of the weekend days. Sunday had the least risk whereas Weekdays had the greatest risk per unit of distance travelled. Among the months November had a relatively greater risk while August had least risk of road accident occurrence per unit of distance travelled. Analysis of the statistical model results revealed that winter and autumn months are associated with more risk in comparison to spring and summer months. The risk per unit of travel on weekdays varies substantially through the year. Greatest risk is associated with weekdays in winter and autumn. Saturdays are comparatively safer than weekdays of the same month but in winter months they have more risk per unit of travel than do some of the weekdays in spring and summer (April and July) months. Sunday carried the lowest risk per unit of travel than all other days and this varied relatively little through the year.

The variables of Christmas, New-Year, and other Public holidays are associated with lower number of road accidents occurring on these days. However, it was not possible to assess risk on these days because no corrections are available for distance travelled on them. The time variable had a negative coefficient which indicates that risk per kilometre of travel declined during the study period of 1991 to 2005. An increase in the distance travelled per vehicle is associated with an increase in the risk of road accident involvement per unit of distance travelled. Travel in police force areas with greater distance travelled per vehicle is associated with a greater risk of a road accident per kilometre. It was also found that police force areas with a greater population density had a greater estimated risk of road accidents per kilometre of travel. However, police forces with a greater number of vehicles per head of population tended to have smaller risks.

The joint use of road accident and meteorological data revealed that higher rainfall was associated with a greater risk of road accidents and police force areas that experienced less

rainfall would have a smaller risk if all other variables remained equal whereas increase in the mean minimum monthly temperature is associated with reduced risk per unit of travel.

Analysis of road accidents using statistical models with joint use of information on road class and vehicle type revealed further associations. Motorways were found to have less risk of accidents per unit of distance travelled for each mode that used them than all other road classes. Urban roads carried the greatest risk of road accidents. Cars had a lower accident risk per kilometre of travel than all other modes. The interaction variables used in the model highlighted the greater risk for car on motorway, pedal cycles on A roads, car on rural minor roads and goods vehicles on minor roads in addition to their main effects. It was also found that for buses the greatest accident risk per unit distance of travel is on urban A roads. Generally pedal cycle and motorcycle are associated with greatest risk per unit of distance travelled than other modes. Leisure motorcycling is associated with greater frequency of involvement in road accidents than other forms of motorcycle usage, though it was not possible to assess risk as no corrections are available for distance travelled. It was also observed that the risk of vehicle involvement in road accidents per unit of distance travelled is high on Monday and Friday in comparison to other days of the week. Similarly September and November have the greatest risk among the month of year.

Cars are involved in injury accidents more frequently than any other kind of vehicle. In view of this, we investigate car casualties in different age and gender groups further by considering their rate of injury in road traffic accidents per million of their population. This can be achieved by analysis of the coefficients in a statistical model of age and gender-specific road accident casualty that has relevant population as offset. Young persons (under 17) and older people (60+) were least likely to be casualties in car accidents. The highest casualty rate per million person-years was found to be in the age range 17 to 20. A clear pattern was observed of decreasing road casualty rate with increasing age. Males had greater casualty rate per million person-years and this was particularly so for young males (aged 17-30).

For transport modes other than the car it was found that pedestrians and motorcyclists between the ages of 17 to 20 have a greater rate of casualty per million person-years than any other age group. Older people (70+) have a greater rate of bus casualty whilst the age range of 21 to 29 had the greatest rate of bicycle casualty. It was also found that, in all cases except travel by bus, the rate per million person-years of road casualty decreased with increasing

age, which could be because people become more experienced over time and safety of their travel behaviour improves. Together, the road accident and casualty data also showed that the greatest rate per million person-year for pedestrian and car casualties occurred in the autumn and winter months while bicyclists' and motorcyclists' greatest rate occurred in the summer months, when their travel activity is likely to be greatest. Friday is associated with greater casualty rate than any other day of the week in each of the modes except bicyclists where Wednesday was found to have the highest rate. Weekdays generally have greater casualty rate than weekends but car travellers have greater casualty rate on Saturday than weekdays except Friday.

6.3 COMPARISON OF STATISTICAL TECHNIQUES USED IN THIS STUDY

The generalized linear model (GLM), generalized estimation equation (GEE) and hierarchical generalized linear models (HGLM) methodologies were all used in this study. It was found that the road accident data is over-dispersed relatively to a Poisson process, as a result of which a negative binomial regression was preferred. A generalized estimation equation (GEE) with autoregressive error terms of order 1 was preferred over the generalized linear model (GLM) because of the presence of serial correlation in the data. It was also found that if the serial correlation was not accommodated then its presence affected the significance levels and in some cases it affected the estimates of the model parameters.

In this particular case it was observed that some of the meteorological effects could be represented through the month when the AR1 error structure was allowed. Deviations from the mean minimum temperature for a month followed a pattern that is represented through the AR1 error term. The HGLM is more computationally demanding than the other techniques but it has the advantages over GLM of joint modelling of mean and dispersion, and it accommodated both within and between category variance among the observations. In this application the HGLM model was preferred over the GEE-AR1 model due to the additional capability it has in incorporating random effects, and in modelling the mean and dispersion jointly. Use of these capabilities is justified by the substantial improvement in model fit that is achieved. Within the HGLM modelling approach, several different variants of likelihood are recommended to be used as objective criteria in estimating the components of an HGLM. The adjusted profile likelihood (APL), which is offered as a criterion for fitting the dispersion model, was found to be unreliable as implemented in GenStat version 12. Consequently the

extended quasi likelihood (EQL) was used instead. It is found that HGLM cannot accommodate time series data due to which the coefficients of some of the variables estimated using it may be unreliable. Due to this, its results need to be interpreted with care and suitable verification checks should be applied.

It was also found that the generalized linear model (GLM) is easy to compute and can accommodate large datasets. On the other hand GEE and HGLM are harder to compute and had difficulty in accommodating large national datasets. The computation time of the GLM with 279,429 observations was 2 to 5 minutes, for the GEE-AR1 model this time was 7 to 10 hours and for the HGLM with 29,216 observations it was 3 to 5 hours, using a Dell Inspiron PC with 2GHz Intel Core Duo, 3GB RAM running under the Windows Vista operating system.

6.4 RISK ESTIMATED FOR VARIOUS GROUPS

The risk per unit of exposure was estimated for each of the vehicle classes, and road classes, whereas casualty rate per million person-years was estimated for each age groups, and gender that were used in this study. In Chapter 4 the risk per unit of travel was estimated for each road and mode combination. These risk values can be used to highlight those combinations that need most attention in reducing road accidents. The results obtained from this study can be used to inform education and promote safer use of road and vehicle combinations. The range of risk per kilometre of travel as shown in Table 4.14 varies substantially by mode of travel and by class of road. It is found that the two groups that have the highest risk per unit of travel are pedal cycles and motorcycles on urban A roads. The other groups that also have high risk per unit distance of travel are motorcycles and goods vehicles on urban minor roads. From the perspective of reducing the number of casualties, the combination of travel by car and urban roads could be prioritised because of the high numbers arising from high distance of travel despite having a lower risk per unit distance travelled than most other groups.

Among the age and gender group combinations that were explored, it was found that young adults (of both genders) aged between 17 and 20 had the greatest road casualty rate per million person years in all vehicle types except buses, whereas older people had a greater rate. Among all combinations of age, gender and vehicle type, car users aged 17 to 20 had the greatest casualty rate per million person years. At all ages, females had a greater casualty rate

per person-year than males while travelling on by bus. Those females under 17 and in the 40 to 59 age group had slightly higher car casualty rate per million person-years than males.

6.5 IMPLICATIONS FOR ROAD SAFETY RESEARCH AND POLICY

The UK Government set out a broad strategy which includes casualty reduction targets mainly by reducing the numbers of casualties up to the year 2010. This was supported by a road safety programme with key objectives to explore the scale and nature of road accidents to identify high risk groups and to understand the travel behaviour. The Department for Transport produces quarterly reports comparing the number of casualties of different groups and by police forces and relates them to a base year and estimate the reduction targets. However, they have not yet highlighted the risk for different groups especially road and vehicle combinations. The results produced by them are widely used by the media, road safety organisations, local authorities and planning organisations to create awareness for improving the road safety. The estimated risk and other results produced in this thesis complement and strengthen those already established. The identification groups with high risk of involvement in road accidents will help individuals on making choices for their journey, and society as whole can benefit from it. Devising a complete road safety policy based on these results is beyond the scope of this thesis but this study highlights some of the important facts:

- There is need to relate the road accident data to the other national data sources. Linking the various sections of STAST19 data to the other sources of information on travel activity is of high importance. Doing that will bring the relative risks of road and vehicle combination, gender and age groups into focus.
- Identification of Weekday and November with greatest risk per unit of travel emphasise the need to focus on travel behaviour during these days. If the high risk in November is mainly due to weather then local authorities should be encouraged to plan special measures for this. However, for individual persons this should serve as a message to be careful while travelling on these days. The high risk per unit of travel on any Weekday might be due to change in either travel pattern or travel behaviour. Realising the sensitivity of the issue, the Department for Transport has already commissioned some of the projects to explore work-related road accidents and contributory factors to them.
- The greater risk per unit of travel in areas with higher population density complements our findings of Chapter 4 that urban roads have greater risk. This research finding lends

support to the government argument for 20 mph zones and home zones to reduce the speed in urban areas.

- The decrease in risk per unit of travel with increase in vehicles per head of population in Chapter 2 shows that affluent areas will have less risk. This research finding further supports the agenda of Department for Transport which led to allocation of the resources to investigate the relationship between the child casualties and social deprivation index.
- The research findings of Chapter 3 show the effect of meteorological factors on the risk per unit of travel. This suggests that those local authorities could use preventive measures such as special sign posts and other engineering measures like increase visibility on the streets to reduce the increased risk due to rain.
- Research finding of Chapter 4 are new and Department for Transport may investigate details of the risk per unit of travel associated with various road and vehicle combinations. Until now the focus of Department for Transport was on comparison of STATS 19 data and Hospital admission data (Hospital Episode Statistics, HES). This research has raised the importance of linking the national datasets. Linking different sections of STATS 19 data will highlight new areas of focus to improve road safety.
- The government is pursuing the policy of modal shift by encouraging people to walking and cycling modes from Cars, even though it is found to be the safest mode. Looking at the risks per unit of travel associated with walking and cycling it is strongly suggested to provide full safety measures on the road before implementing any wide-spread programme.
- The research findings of Chapter 5 also support and strengthen the agenda of Department for Transport for reducing the child and old age person's casualties. However, new insights have been gained by combining the casualty information with information from accident section of STATS 19 data (when and where they occur). The disaggregation of results by gender highlights the importance of focusing on specific gender, age groups and mode in any subsequent road safety plans.

On the methodological development, various statistical techniques were compared. It is found that generalized linear model (GLM) is generally adequate to model large datasets, it can accommodate over-dispersion but not serial correlation. The coefficients and significance levels of some variables were found to change substantially if the presence of serial correlation is not respected. Generalized estimation equation (GEE) is computationally demanding but it can accommodate the serial correlation that was found to be present in the data, which GLM cannot accommodate. Hierarchical generalized linear Model (HGLM), which is relatively new, is more

computationally demanding than either GLM or GEE. It has additional benefits of using random terms and dispersion modelling. However, HGLM cannot accommodate serial correlation, and due to this its results should be interpreted with care and suitable verification checks applied before initiating road safety programmes on the basis of results from it.

6.6 FUTURE WORK

In this study, the research opportunity has been taken to use national road accident datasets to model the number of road accidents occurring on each day, number of vehicles involved in road accidents on each day, and number of road casualties occurring each day. This work could be further developed in four directions:

- The methodology developed to model road accidents from national datasets can be used at the national level in other countries, especially in developing countries where road accident rates are currently increasing. The road accident datasets will also be compared and the lessons learnt from this study will be applied for the improvement of road safety.
- Further research is required to explore suitable statistical modelling methods to accommodate the spatial autocorrelation among data from the police forces.
- The statistical modelling methods, especially HGLM, applied successfully to count data are not fully mature. Investigations will be required to determine the reasons for their shortcomings with large datasets and to establish appropriate procedures. Further development will be required to enable them to accommodate the spatial and temporal correlation among the data.
- The new data from Motorway Incident Detection and Automatic Signalling (MIDAS) for the western quadrant of the M25 between the junctions 10 (A3) and 15 (M4) includes minute by minute loop detector data about the traffic flow by vehicle class. The information about road accidents by time of the day is also available from the STATS 19 data. Combined use of STATS 19 data and MIDAS data will lead to estimates of the number of vehicles involved in accidents disaggregated by time and vehicle class. This will lead to further insights about the risks per unit of travel for different types of vehicle at different times of day disaggregated by day of the week.

REFERENCES

Abdel-Aty, MA and Radwan AE (2000) Modelling traffic accident occurrence and involvement. *Accident Analysis and Prevention,* 32 (5) 633-642.

Aitkin, M, Anderson, D, Francis, B and Hinde, J (1989) Statistical modelling in GLIM. Oxford university press, Oxford, 217-223.

Amis, G (1996) An application of generalized linear modelling to the analysis of traffic accidents. *Traffic Engineering Control,* 37 (12), 691-696.

Andreescu, PM and Frost, BD (1998) Weather and traffic accidents in Montreal, Canada. *Climate Research,* 9, 225-230.

Andrey, J and Yagar, S (1993) A temporal analysis of rain-related crash risk. *Accident Analysis and Prevention,* 25 (4), 465-472.

Andret, J and Olley, R (1990) Relationship between weather and road safety: Past and future research directions. Climatological Bulletin, 24 (3), 123-127.

BBC Weather Centre, accesses on 25th July 2010, www.bbc.co.uk/ukweather/year_review/review/december2004_review.shtml.

Behr, JB (1963) Research on road safety. Her majesty's stationary office, UK.

Bertness, J (1980) Rain related impacts on selected transportation activities and utility services in the Chicago area. Journal of Applied Metrology, 19 (5), 545-556.

Palutikof, JP (1991) Road accident and weather. *Highway Meteorology (editors Perry AH and Symons LJ) 163-189, Taylor and Francis Group* an imprint of Chapman and Hall, London. (R)

Bester, CJ (2001) Explaining national road fatalities. *Accident Analysis and Prevention,* 33 (5) 663-672.

Bird, RN, Wedagama, DMP and Metcalfe AV (2006) The influence of urban land-use on non-motorized transport casualties. *Accident Analysis and Prevention,* 38 (6) 1049-1057.

Box, GEP and Cox DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 26, 211-52.

Boyanowski, EO, Calvert, J, Young, J and Brideau, L (1981) Toward a thermoregulatory mode of violence. Journal of Environmental Systems, 11 (1), 81-87.

Brijs, T, Karlis, D and Wets, G (2008) Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention,* 40 (3) 1180-1190.

Brotsky, H and Hakkert, AS (1998) Risk of a road accident in rainy weather. *Accident Analysis and Prevention,* 20(3) 161-176.

Campbell, ME (1971) The wet pavement accident problem: breaking through. *Traffic Quarterly,* 25, 209-214.

Chandler, RE and Bate S, (2007) Inference for clustered data using the independence loglikelihood. *Biometrika,* 94 (1), 167-183.

Chandler, RE and Scott, M, (2011) Statistical methods for trend detection and analysis, Wiley Series.

Chatterjee S and Hadi AS, (2006) Regression analysis by example, Fourth edition, Wiley Interscience.

Codling, PJ (1974) Weather and road accidents. *Climatic resources and economic activity.* (JA Taylor, ed) 205-222. Newton Abbot, UK: David and Charles Holdings

Crosta, PM and Packman, IG (2005) Faculty productivity in supervising doctoral students dissertations at Cornell university. *Economics of Education Review,* 24 (1), 55-65.

de Freitas, CR (1975) Estimation of the disruptive impact of snowfalls in urban areas. *Journal of Applied Metrology,* 14(6), 1166-1173.

Department for Transport, Highways economic note no. 1, 2005 valuation of the benefits of prevention of road accidents and casualties, 2007, retrieved April 2007, http://www.dft.gov.uk/pgr/roadsafety/ea/.

Department for Transport, Focus on personal travel, 2001, retrieved March 2005, http://www.dft.gov.uk/pgr/statistics/datatablespublications/personal/focuspt/2001/.

Department for Transport (2010) Reported road casualties Great Britain: Annual report 2009.

Department for Transport (2011): Reported Road Casualties in Great Britain: Annual report 2010.

Department for Transport (2011): National Travel Survey 2010. http://www.dft.gov.uk/site/archive.

Department for Transport, Road Length Statistics, retrieved 2nd February 2011, http://www.dft.gov.uk/statistics/series/road-lengths/.

Draper, NR and Smith, H (1998) *Applied Regression Analy*sis, 3rd edition, Wiley Series.

Edwards, JB (1993) The influence of weather on road accidents in England and Wales. PhD Thesis, Cardiff University.

Edwards, JB (1998) The relationship between road accident severity and recorded weather. *Journal of Safety Research,* 29 (4), 249-262.

Edwards, JB (1996) Weather-related road accidents in England and Wales: a spatial analysis. *Journal of Transport Geography*, 4 (3), 201-212.

Eisenberg, D (2004) The mixed effects of precipitation on traffic crashes. *Accidents Analysis and Prevention*, 36 (4), 637-647.

Elvik, R and Vaa, T (2004) *The handbook of road safety measures,* Elsevier, Oxford, U.K., 29-93.

Evans, L (2004) *Traffic safety*. Science Serving Society, Michigan, USA.

Fontaine H and Gourlet, Y (1997) Fatal pedestrian accidents in France: A typological analysis. *Accidents Analysis and Prevention*, 29 (3), 303-312.

Francis, B, Green, M and Payne, C (1993) The GLIM system. New York: Oxford University Press.

Fridstrom, L, Ifver, J, Kulama, R and Thomsen, LK (1995) Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accident counts. *Accidents Analysis and Prevention*, 27 (1), 1-20.

Grafen, A and Hails, R (2002) *Modern statistics for the life sciences*. Oxford university press, United Kingdom.

Greibe, P (2003) Accident prediction models for urban roads. *Accidents Analysis and Prevention*, 35 (2), 273-285.

Gujarati DN and Porter DC, (2009) *Basic Econometrics*. McGraw-Hill International Edition.

Hadi, MA, Aruldhas, J, Chow, L-F. and Wattleworth, JA (1995) Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record: Journal of the Transportation Research Board*, 1500, 169-177.

Hardin, J and Hilbe, J (2001) Generalized linear models and extensions. Stata Press, Texas, USA.

Hardin, J and Hilbe, J (2003) Generalized estimation equations. Chapman and Hall/CRC,

Haghighi-Talab, D (1973) An investigation into the relationship between rainfall and road accident frequencies in two cities. *Accidents Analysis and Prevention*, 5 (4), 343-349.

Hall, RD (1986) Accidents at 4-arm urban traffic signals. TRRL report 65, Transport and road research laboratory, United Kingdom.

Hijar, M, Carrillo, C, Flores, M, Anaya, R and Lopez, V (2000) Risky factors in highway traffic accidents: a case control study. *Accidents Analysis and Prevention*, 32 (5), 703-709.

Hilbe, J (1993) Generalized linear models. Stata Technical Bulletin Reprints, 149-159. College Station, TX: Stata Press.

Hilbe, J (2007) Negative Binomial Regression. Cambridge University Press, Cambridge, UK.

Hutchings, CB, Knight, S, Reading, JC (2003) The use of generalized estimation equation in the analysis of motor vehicle crash data. *Accidents Analysis and Prevention*, 32 (2003), 3-8.

Huddart, K, Lawson, G, Purcell, R, Robinson, R, Sabey, B, Stewart, D, Tranter, S and Raikes, T (1991) *Towards safer roads in developing countries: A guideline for planners and engineers.* Transport and road research laboratory, United Kingdom.

International traffic safety data and analysis group, Downward trend in road traffic fatalities in IRTAD member countries in 2008, retrieved 15th August 2010, http://cemt.org/IRTAD/IRTADPublic/index.htm.

Jacobs, G, Aeron-Thomas, A and Astrop A (2000) Estimating global road fatalities. Crowthorne, Transport research laboratory, TRL report 445.

Jones, B, Janssen, L and Mannering, F (1991) Analysis of frequency and duration of freeway accidents in Seattle. *Accidents Analysis and Prevention*, 23 (4), 239-255.

Joshua, SC and Garber NJ (1990) Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation planning and Technology,* 15 (1), 41-58.

Jovanis, PP and Chang, H-L (1986) Modelling the relationship of accidents to miles travelled. *Transportation Research Record: Journal of the Transportation Research Board,* 1068, 42-51.

Kaell, MD (1995) Pedestrian exposure to risk of road accident in New Zealand. *Accidents Analysis and Prevention*, 27 (5), 729-740.

Keay, K and Simmonds, I (2005) The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accidents Analysis and Prevention*, 37(1), 109-124.

Kendall, M and Ord, JK (1990) Time series. New York: Edward Arnold.

Kulmala, R (1995) Safety at rural three-and four-arm junctions: Development and application of accident prediction models. Doctor of technology thesis, Helsinki University of Technology, Finland.

Kutner, M and Neter J (2004) Applied linear regression models. McGraw-Hill Irwin.

Ladron de Guevara, F, Washington, SP and Oh, J (2004) Forecasting crashes at the planning level simultaneous negative binomial crash model applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board*, 1897, 491-499.

Lee, Y and Nelder JA (1996) Hierarchical generalized linear models (with discussion). Journal of Royal Statistical Society, Series B, 58, 619-678.

Lee, Y (2004) Estimating intraclass correlation for binary data using extended quasi-likelihood. *Statistical Modelling*, 4 (2), 113-126

Lee, Y and Nelder JA (2001) Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1 (1), 3-16.

Lee, Y and Nelder JA (2001) Hierarchical generalized linear model: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrica*, 88 (4), 987-1006.

Lee, Y, Nelder JA and Noh, M (2007) H-likelihood: Problems and solutions. *Statistics and Computing*, 17 (1), 49-55.

Lee, Y, Nelder JA and Pawitan, Y (2006) Generalized linear models with random effects: unified analysis via h-likelihood. Chapman and Hall/CRC,

Legge, M, Ryan, GA and Rosman, DA (1998) Age related changes in drivers' crash risk and crash type. *Accident Analysis and Prevention*, 30 (3), 379-389.

Leveine, N, Lim, KE and Nitz LH (1995) Daily fluctuations in Honolulu motor vehicle accidents. *Accident Analysis and Prevention*, 27 (6) 785-796.

Liang, KY and Zeger SL (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrica*, 73 (1), 13-22.

Lord, D and Persaud BN (2000) Accident prediction models with and without trend: Application of the generalized estimation equations procedure. *Transportation Research Record: Journal of the Transportation Research Board*, 1717, 102-108.

Maher, MJ and Summersgill, I (1996) Comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention*, 28 (3), 281-296.

Maycock, G and Hall, RD (1984) Accidents at 4-arm roundabouts. TRRL report 1120, Transport and road research laboratory, United Kingdom.

McCarthy, P.S., Public policy and alcohol related crashes among old driver, www.econ.gatech.edu/papers/mccarthy_CADTS_paper_SOEwebst_091402.pdf. Accessed March, 2005.

McConway, KJ, Jones, MC and Taylor PC (1999) *Statistical modeling using GenStat*. Arnold, United Kingdom.

McCullagh, P and Nelder JA (1983) *Generalized Linear Models*, Chapman and Hall, USA.

Memon, AQ (2006) Road accident prediction models developed from national database: Poisson and negative binomial regressions. *85th Meeting of Transportation Research Board*. Washington, DC.

Memon, A Q (2008) Comparison of generalised linear model and generalised estimation equation for modelling road accidents from national datasets. 11th world conference on Transport Research. University of California, Berkley.

Meteorological office UK, Historic Station Data, retrieved 5th January 2011, http://www.metoffice.gov.uk/climate/uk/stationdata/.

Miaou, SP and Lum, H. (1993) Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis and Prevention*, 25(6), 689-709.

Miaou, SP (1994) Relationship between truck accidents and geometric design of road sections: Poisson and negative binomial regressions. *Accident Analysis and Prevention*, 26 (4), 471-482.

Montgomery, D and Runger, G (2010) Applied statistics and probability for engineers. John Wiley and Sons, Inc.

Nelder, JA and Pregibon, D (1987) An extended quasi-likelihood function. *Biometrika*, 74 (2), 221-232.

Noland, RB and Quddus MA (2005) Congestion and safety: A spatial analysis of London. *Transportation Research Part A*, 39 (7-9), 737-754.

Office for National Statistics UK, Population estimates for England and Wales, Scotland and Northern Ireland, Population Estimates, Time series, 1971 to current year, assessed on 3rd March 2011, http://www.statistics.gov.uk/hub/index.html

Ogden, KW (1996) *Safer roads: A guide to road safety engineering*. Avebury Technical, England, 1-93.

Salifu, M (2004) Accident prediction models for unsignalised urban junction in Ghana. Journal of the International Association of Traffic and Safety Sciences, 28 (1), 68-81.

Satterthwaite, SP (1976) An assessment of seasonal and weather effects on the frequency of road accidents in California. *Accident Analysis and Prevention*, 8 (2), 87-96.

Ida van Schalkwyk, Sudeshma, M and Washington, S (2006) Incorporating weather into region-wide safety planning prediction models. *85th Meeting of Transportation Research Board*. Washington, DC.

Safety support, 'esafety activities: National activities'. [Online], 20 September 2007. http://www.escope.info/en/welcome.htm.

Schwarz, Gideon E (1978) Estimating the dimension of a model. Annals of Statistics 6(2): 461-464.

Shankar, V, Mannering, F and Barfield, W (1995) Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention*, 27(3), 371-389.

Shankar, V, Milton J and Mannering, F (1997) Modelling accident frequencies as zero-altered probability process: an empirical analysis. *Accident Analysis and Prevention*, 29(6), 829-837.

Sittikariya, S and Shankar, V. (2005) Accounting for serial correlation in count models of traffic safety. *Journal of the East Asia Society for Transport Studies*, 6, 3645-3657.

Smith, K (1982) How seasonal and weather conditions influence road accidents in Glasgow. *Scottish Geographical Magazine,* 98, 103-114.

STATA Press (1985), Reference manual *Longitudinal/ Panel data*, Release 9, USA.

STATA Press (2001), Reference manual *Generalized Linear Models and Extensions.* USA.

Stern, E and Zehavi, Y (1990) Road safety and hot weather: A study in applied transport geography. *Transactions of the Institute of British Geographers,* New series, 15 (1), 102-111.

Tanner, JC (1953) Accidents at rural three-way junctions. *Journal of Institution of Highway Engineers* (11), 56-57

UCLA (2009), Academic technology services, Statistical consulting group, accesses on 30th March 2009. www.ats.ucla.edu/stat/stata/default.htm

Ulfarsson, GF and Shankar VN (2003) An accident count model based on multi-year cross-sectional roadway data with serial correlation. *Transportation Research Record: Journal of the Transportation Research Board,* 1840, 193-197.

Umer, RS, Mackay, MG and Hills, BL (1996) Modelling the conspicuity-related motorcycle accidents in Seremban and Shah Alam, Malaysia. *Accident Analysis and Prevention*, 28 (3), 325-332.

Van den Bossche, Filip, Wets, Geert and Brijs, Tom (2006) Predicting road crashes using calendar data. *85th Meeting of Transportation Research Board.* Washington, DC.

Wang, X. and Abdel-Aty, MA (2006) Crash estimation at signalized intersections along corridors: Analyzing Spatial Effect and Identifying Significant Factors, *85th Meeting of Transportation Research Board*. Washington, DC.

Washington, PS, Karlaftis, MG and Mannering, FL (2003) *Statistical and econometric methods for Transportation data analysis*. Chapman and Hall/CRC, USA.

Wedderburn, RWM (1974) Quasi-likelihood functions, generalize linear models and the Gauss-Newton method. *Biometrica*, 61 (3), 439-447.

Weiner, JS and Hutchinson, JCD (1945) Hot humid environment, its effect on the performance of a motor coordination test. *British Journal of Industrial Medicine*, 2 (3) 154-157.

World Health Organization, World report on road traffic injury prevention, ISBN no. 9241562609, 2004, retrieved 6th June 2006, http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/index.html

Zhang, J, Lindsay, J, Clarke, K, Robbins, G and Mao, Y (2000) Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. *Accident Analysis and Prevention*, 32 (1), 117-125.

APPENDIX

APPENDIX A1.1

1. Road accident reporting system in some of OECD countries:

Road accident statistics are used for evaluating the level of road safety at both national and international levels. In addition, data is used for road safety research, identifying accident hot spots and estimating road safety risks. In various countries road accidents are mostly reported by the police. More industrialised counties often have systematic procedures from recording and coding to the management of their road accident databases (Safety support, 2007). The procedure adopted by some of OECD countries is given in detail below:

1.1 Finland

In Finland, police collect the road accident data as part of their routine police activity. Statistics Finland receives data from police that is entered into the PATJA information system of police affairs. Statistics Finland is responsible for the maintenance of the database and it also controls access to data. Statistics Finland makes further checks and then supplements the data by comparing it with other datasets including causes of death, national road administration data on accident locations and casualties. Data can also be acquired in files from Statistics Finland whereas monthly statistics are made available for users on the website of Traffic Safety of Finland.

1.2 France

In France the national accident database is derived from police force reports. The road accident data is based on BAAC forms (Bulletin d'Analyse des Accidents Corporels de la Circulation). These standard electronic forms are filled in for every traffic accident resulting in personal injury. The police forces send the BAAC files to the National Inter-Ministerial Road Safety Observatory (ONISR). After transmitting these files to ONISR, these are further checked by Service d'Etudes Techniques des Routes et Autoroutes (SETRA) which is a technical service of the Ministry of Infrastructure. SETRA controls the quality, identifies duplicate entries, cross-checks the BAAC against the local road safety figures in order to identify any missing data and finally compiles the files on a monthly basis. Database

maintenance is responsibility of SETRA. ONSIR controls access to data and publishes official road safety statistics based on the BAAC data.

1.3. Germany

For every road accident in Germany the police are informed and investigate the accident scene. This investigation results in the production of a standardised computer police report. The data collected in the report includes information about vehicles, weather, persons involved, accident scene, and circumstances of the accident. Some major cities have a special police unit (Verkahsunfalldienst VUD), which deals with traffic accidents. According to law, only accidents involving vehicular traffic are recorded therefore accidents involving only pedestrians are not included in the federal statistics. Since 1975, accident causes have also been registered. From the police stations, data is transmitted to the federal office of statistics, Statistisches Bundesamt (StBA) who own and maintain the database. The data is used by the federal government for monitoring the development of road traffic and for policy making. Aggregated data are made available to members of public and can be accessed over the internet.

1.4. Italy

In Italy, the national road accident database is maintained by National Institute of Statistics (ISTAT). Annual statistics on road accidents are published by ISTAT to inform members of the public about road safety issues. The accident database contains information about all traffic accidents and injuries. In Italy, data collection is not standardised as police reports are drafted according to local protocols but all reports contain data concerning the vehicle, environment, weather conditions, and accident description. All the data gathered by the police is then used to complete the ISTAT module which is a standardised form. This form is then sent to the Provincial Capital Statistical office and thence to ISTAT. Data are used by ISTAT for the production of the official statistics. These publications promote the development of research activities. Data become available yearly and are free of charge to research institutes.

1.5. Netherlands

In the Netherlands, accident data are collected by the police who complete accident reporting forms at the scene. These forms are sent to a central coding agency which incorporates the data into a database. Data are coded according to special guidelines and weekly checks are made. This data is linked with GIS maps and in some cases it is also linked with licence plate registration, and other databases to improve data quality. The data are owned by the Ministry of Transportation. Data can usually be purchased but a few research organizations have access to the aggregated data.

APPENDIX A1.2

Road safety plans of some of OECD countries:

The road safety plans of Denmark, Netherlands, Sweden, Finland, Canada, and United Kingdom are as follows:

1. Danish road safety plans

The vision and central theme of Danish Road Safety Strategy is embedded in their slogan 'Every Accident is One too Many'. The vision was launched in the Danish Government's action plan on road safety. The vision sets a course towards a future road system that is without any road accident whatsoever and therefore retains a focus on preventive measures. Although the vision is to prevent all road accidents, the road safety policy objective for 2012 is:

- To reduce the number of fatal and serious injuries by at least 40 percent in 2012 compared to the baseline average of 1998. This requires that the total number of fatalities and of seriously injured casualties should not exceed 300 and 2,443 respectively.
- The main areas of focus for road safety are speeding, alcohol, cyclists, and junctions, which together are factors in almost 85 percent of road accidents.

The Danish national commission on road safety is responsible for evaluating the national road safety plan. The national plan is monitored three times per year and evaluated after every four years. Monitoring indicators for road safety in Denmark are the number of road accidents, fatalities and serious injuries, accidents at intersections, accidents involving cyclists, speed of vehicles involved in road accidents, and drink-driving.

2. Netherlands road safety plans

The Dutch road safety policy relies on the concept of sustainable road safety. In the 1980s, the Dutch Ministry of Transport, Public Works and Water Management set road safety targets to reduce annual fatalities by 50 percent and to have 40 percent fewer hospital admissions by 2010 compared to 1986.

The Dutch government considers speed enforcement, speed management, and the roadworthiness test for vehicles as areas of focus for road safety. As a result of this, many infrastructural traffic calming measures have been adopted including chicanes, speed humps, and roundabouts. Speed cameras have also been used on motorways, on secondary roads and in urban areas. The Dutch Ministry of Transport, Provinces and Municipalities is responsible for monitoring and evaluating the road safety plans. At national level, road safety policy effect reports are released every year with a comprehensive report every four years. Monitoring indicators are the number of accidents, fatalities and casualties at all severity levels. Risk exposure and seatbelt use are also monitored.

3. Swedish road safety plans

The central theme of the Swedish road safety plan is the Vision Zero concept. Fatalities and serious casualties are regarded unacceptable in Sweden. In 1977, the Swedish Parliament approved the Vision Zero programme which states that 'Nobody should be killed or seriously injured within the road transport system'. According to this, road transport systems structure and function should be brought into line with the demands that this goal entails. The Vision Zero principles are:

- The traffic system should be adapted to take better account of the needs, mistakes, and vulnerabilities of road users;
- The level of violence that the human body can tolerate without sustaining fatal or seriously injury forms the basic parameter in the design of the road transport system; and
- Vehicle speed is the most important regulating factor for safe road traffic. It should be determined by the technical standards of both roads and vehicles so as not to exceed the level of violence that human body can tolerate.

The Swedish target for the year 2000 was to reduce fatalities by 25 percent compared with a base year of 1996 whereas the target for 2007 was a reduction of fatalities by 50 percent. Various priorities including safer traffic in built-up areas, safer vehicles, cable guardrails, safer motorways, seatbelt reminders, cycle helmets, and safer commercial vehicles in operation were considered. The Swedish Road Administration and the Swedish National

Road and Transport Institute are responsible for evaluating road safety plans. Accident, fatality and casualty prediction curves are produced and monitored regularly.

4. Finnish road safety plans

Finland is one of the top countries in world for road safety due to its high standards and performance. The Finnish National Road Safety Plan 2005 was established by the Consultative Committee on Road Safety, under the responsibility of the Ministry of Transport and Communication. Various specific objectives and activities were defined for 2005. The base year used for the number of fatalities is 1989 (with 734 fatalities). The target set for 2000, was to reduce fatalities by 50 percent (367 fatalities). The actual number of fatalities in 2000 at 396 exceeded this by 29. The target set for 2005 was to reduce fatalities by 65 percent to fewer than 250 fatalities. Since the target for 2000 was not met, the Consultative Committee on Road Safety presented an updated road safety programme for 2001-2005, containing more intensified and more effective road safety measures. The vision of the Finnish government about road safety is based on the principle that the road transport system should be designed so that nobody should die or be seriously injured on its roads. The aim of the 2001-2005 road safety programme was to create the conditions for a continuous improvement in the transport system, with the target of no more than 100 traffic fatalities by 2025. The priority method that has been identified to improve road safety in Finland is to curb traffic growth with the aim of reducing the likelihood of accidents by influencing choice of mode of transport and effective use of technology.

5. Canadian road safety plans

The Canadian government has an ambitious target to achieve a 30 percent decrease in the number of people fatally or seriously injured on its roads by 2010 from the 1996 baseline. The government's vision is to have the safest roads in world and it intends to achieve this by implementing high quality data collection systems, the dedicated application of problem solving, partnership building, enforcement, education, and evaluation of the programme. Within the main targets are the following objectives, to:

- achieve a 95 percent minimum seatbelt wearing rate and proper use of child restraints;

- reduce the number of road users fatally or seriously injured on rural roads by 40 percent;
- reduce the number of fatally or seriously injured casualties involving drinking and driving by 40 percent;
- reduce the number of drivers killed or seriously injured in speed and intersection related accidents by 20 percent;
- reduce the number of young drivers/riders (of 16 to 19 years) killed or seriously injured in accidents by 20 percent;
- reduce the number of people killed or seriously injured in accidents involving commercial carriers by 20 percent; and
- reduce the number of vulnerable road users (pedestrians, motorcyclists, and cyclists) killed or seriously injured by 30 percent.

Appendix Table A2.1: Days coded as public holidays, Christmas holidays and New-year holidays

H	N	C	1991	Day of week	Holiday event	H	N	C	1995	Day of week	Holiday event
√	√	X	1st January	Tuesday	New Year's day	√	√	X	1st January	Sunday	New Year's day
√	X	X	29th March	Friday	Good Friday	√	√	X	2nd January	Monday	New Year's day holiday
√	X	X	31st March	Sunday	Easter Sunday	√	X	X	14th April	Friday	Good Friday
√	X	X	1st April	Monday	Easter Monday	√	X	X	16th April	Sunday	Easter Sunday
√	X	X	6th May	Monday	Early May bank holiday	√	X	X	17th April	Monday	Easter Monday
√	X	X	27th May	Monday	Bank holiday	√	X	X	8th May	Monday	Early May bank holiday
√	X	X	26th August	Monday	Summer bank holiday	√	X	X	29th May	Monday	Spring bank holiday
√	X	√	25th December	Wednesday	Christmas day	√	X	X	28th August	Monday	Summer bank holiday
√	X	√	26th December	Thursday	Boxing day	√	X	√	25th December	Monday	Christmas day
						√	X	√	26th December	Tuesday	Boxing day
H	**N**	**C**	**1992**	**Day of week**	**Holiday event**	**H**	**N**	**C**	**1996**	**Day of week**	**Holiday event**
√	√	X	1st January	Wednesday	New Year's day	√	√	X	1st January	Monday	New Year's day
√	X	X	17th April	Friday	Good Friday	√	X	X	5th April	Friday	Good Friday
√	X	X	19th April	Sunday	Easter Sunday	√	X	X	7th April	Sunday	Easter Sunday
√	X	X	20th April	Monday	Easter Monday	√	X	X	8th April	Monday	Easter Monday
√	X	X	4th May	Monday	Early May bank holiday	√	X	X	6th May	Monday	Early May bank holiday
√	X	X	25th May	Monday	Spring bank holiday	√	X	X	27th May	Monday	Spring bank holiday
√	X	X	31st August	Monday	Summer bank holiday	√	X	X	26th August	Monday	Summer bank holiday
√	X	√	25th December	Friday	Christmas day	√	X	√	25th December	Wednesday	Christmas day
√	X	√	26th December	Saturday	Boxing day	√	X	√	26th December	Thursday	Boxing day
√	X	√	28th December	Monday	Bank holiday						
H	**N**	**C**	**1993**	**Day of week**	**Holiday event**	**H**	**N**	**C**	**1997**	**Day of week**	**Holiday event**
√	√	X	1st January	Friday	New Year's day	√	√	X	1st January	Wednesday	New Year's day
√√	X	X	9th April	Friday	Good Friday	√	X	X	28th March	Friday	Good Friday
√	X	X	11th April	Sunday	Easter Sunday	√	X	X	30th March	Sunday	Easter Sunday
√	X	X	12th April	Monday	Easter Monday	√	X	X	5th May	Monday	Easter Monday
√	X	X	3rd May	Monday	Early May bank holiday	√	X	X	26th May	Monday	Spring bank holiday
√	X	X	31st May	Monday	Spring bank holiday	√	X	X	25th August	Monday	Summer bank holiday
√	X	X	30th August	Monday	Summer bank holiday	√	X	√	25th December	Thursday	Christmas day
√	X	√	25th December	Saturday	Christmas day	√	X	√	26th December	Friday	Boxing day
√	X	√	26th December	Sunday	Boxing day						
√	X	√	27th December	Monday	Bank holiday						
√	X	√	28th December	Tuesday	Bank holiday						
H	**N**	**C**	**1994**	**Day of week**	**Holiday event**	**H**	**N**	**C**	**1998**	**Day of week**	**Holiday event**
√	√	X	1st January	Saturday	New Year's day	√	√	X	1st January	Thursday	New Year's day
√	√	X	3rd January	Monday	New Year's day holiday	√	X	X	10th April	Friday	Good Friday
√	X	X	1st April	Friday	Good Friday	√	X	X	12th April	Sunday	Easter Sunday
√	X	X	3rd April	Sunday	Easter Sunday	√	X	X	13th April	Monday	Easter Monday
√	X	X	4th April	Monday	Easter Monday	√	X	X	4th May	Monday	Early May bank holiday
√	X	X	2nd May	Monday	Early May bank holiday	√	X	X	25th May	Monday	Spring bank holiday
√	X	X	30th May	Monday	Spring bank holiday	√	X	X	31st August	Monday	Summer bank holiday
√	X	√	25th December	Sunday	Christmas day	√	X	√	25th December	Friday	Christmas day
√	X	√	26th December	Monday	Boxing day	√	X	√	26th December	Saturday	Boxing day
√	X	√	27th December	Tuesday	Bank holiday	√	X	√	28th December	Monday	Bank holiday

Appendix Table A2.1: Days coded as public holidays, Christmas holidays and New-year holidays

H	N	C	1999	Day of week	Holiday event
√	√	X	1st January	Friday	New Year's day
√	X	X	2nd April	Friday	Good Friday
√	X	X	4th April	Sunday	Easter Sunday
√	X	X	5thApril	Monday	Easter Monday
√	X	X	3rd May	Monday	Early May bank holiday
√	X	X	31st May	Monday	Spring bank holiday
√	X	X	30th August	Monday	Summer bank holiday
√	X	√	25th December	Saturday	Christmas day
√	X	√	26th December	Sunday	Boxing day
√	X	√	27th December	Monday	Bank holiday
√	X	√	28th December	Tuesday	Bank holiday

H	N	C	2000	Day of week	Holiday event
√	√	X	1st January	Saturday	New Year's day
√	X	X	3rd January	Monday	New Year's day holiday
√	X	X	21st April	Friday	Good Friday
√	X	X	23rd April	Sunday	Easter Sunday
√	X	X	24th April	Monday	Easter Monday
√	X	X	1st May	Monday	Early May bank holiday
√	X	X	29th May	Monday	Spring bank holiday
√	X	X	28th August	Monday	Summer bank holiday
√	X	√	25th December	Monday	Christmas day
√	X	√	26th December	Tuesday	Boxing day

H	N	C	2001	Day of week	Holiday event
√	√	X	1st January	Monday	New Year's day
√	X	X	13th April	Friday	Good Friday
√	X	X	15th April	Sunday	Easter Sunday
√	X	X	16th April	Monday	Easter Monday
√	X	X	7th May	Monday	Early May bank holiday
√	X	X	28th May	Monday	Spring bank holiday
√	X	X	27th August	Monday	Summer bank holiday
√	X	√	25th December	Tuesday	Christmas day
√	X	√	26th December	Wednesday	Boxing day

H	N	C	2002	Day of week	Holiday event
√	√	X	1st January	Tuesday	New Year's day
√	X	X	29th March	Friday	Good Friday
√	X	X	31st March	Sunday	Easter Sunday
√	X	X	1st April	Monday	Easter Monday
√	X	X	6th May	Monday	Early May bank holiday
√	X	X	3rd June	Monday	Golden Jubilee holiday
√	X	X	4th June	Tuesday	Spring bank holiday
√	X	X	26th August	Monday	Summer bank holiday
√	X	√	25th December	Wednesday	Christmas day
√	X	√	26th December	Thursday	Boxing day

H	N	C	2003	Day of week	Holiday event
√	√	X	1st January	Wednesday	New Year's day
√	X	X	18th April	Friday	Good Friday
√	X	X	20th April	Sunday	Easter Sunday
√	X	X	21st April	Monday	Easter Monday
√	X	X	5th May	Monday	Early May bank holiday
√	X	X	26th May	Monday	Spring bank holiday
√	X	X	25th August	Monday	Summer bank holiday
√	X	√	25th December	Thursday	Christmas day
√	X	√	26th December	Friday	Boxing day

H	N	C	2004	Day of week	Holiday event
√	√	X	1st January	Thursday	New Year's day
√	X	X	9th April	Friday	Good Friday
√	X	X	11th April	Sunday	Easter Sunday
√	X	X	12th April	Monday	Easter Monday
√	X	X	3rd May	Monday	Early May bank holiday
√	X	X	31st May	Monday	Spring bank holiday
√	X	X	30th August	Monday	Summer bank holiday
√	X	√	25th December	Saturday	Christmas day
√	X	√	26th December	Sunday	Boxing day
√	X	√	27th December	Monday	Bank holiday
√	X	√	28th December	Tuesday	Bank holiday

H	N	C	2005	Day of week	Holiday event
√	√	X	1st January	Saturday	New Year's day
√	√	X X	3rd January	Monday	New Year's day holiday
√	X	X	25th March	Friday	Good Friday
√	X	X	27th March	Sunday	Easter Sunday
√	X	X	28th March	Monday	Easter Monday
√	X	X	2nd May	Monday	Early May bank holiday
√	X	X	30th May	Monday	Spring bank holiday
√	X	X	29th August	Monday	Summer bank holiday
√	X	√	25th December	Sunday	Boxing day
√	X	√	26th December	Monday	Bank holiday
√	X	√	27th December	Tuesday	Bank holiday

Appendix Table A2.2: Results of models for the 51 police force areas of Great Britain with Ln (Total distance travelled nationally on each day as offset-Dataset 2)

Model	D.F	Scale	Likelihood	*BIC*
1	1	0.74232	-987,439	1,974,891
2	7	0.73186	-985,445	1,970,978
3	12	0.74645	-988,229	1,976,609
4	3	0.73346	-985,742	1,971,521
5	4	0.75254	-989,358	1,978,766
6	18	0.73636	-986,321	1,972,867
7	6	0.74391	-987,729	1,975,533
8	84	0.73539	-986,135	1,973,323
9	12	0.74321	-987,595	1,975,340
10	22	0.73723	-986,477	1,973,230
11	23	0.72661	-984,461	1,969,210
12	24	0.72422	-984,001	1,968,303
13	25	0.72331	-983,823	1,967,959
14	26	0.72319	-983,799	1,967,924
15	76	0.05901	-886,157	1,773,267
16	27	0.40507	-915,701	1,831,741
17	27	0.70358	-980,129	1,960,597
18	27	0.39802	-911,537	1,823,412
19	27	0.40335	-914,953	1,830,244
20	27	0.07353	-995,918	1,992,174
21	42	0.25380	-863,753	1,728,032
22	28	0.40368	-915,390	1,831,130
23	28	0.36379	-902,087	1,804,526
24	28	0.40304	-914,837	1,830,026
25	28	0.37117	-906,009	1,812,370
26	43	0.22768	-855,069	1,710,676
27	29	0.35419	-900,575	1,801,514
28	29	0.35731	-901,171	1,802,705
29	44	0.22756	-855,043	1,710,639
30	30	0.35217	-899,954	1,800,284
31	45	0.21961	-852,461	1,705,485
32	46	0.21924	-852,358	1,705,293
33	96	0.05753	-755,179	1,511,561

D.F= degrees of freedom; BIC= Bayesian information criterion

Appendix Table A2.3: Distance travelled by day of week (index): Great Britain, 1991-2005
Source of data: Department for Transport

Distance travelled by day of week (Index): Great Britain

Year	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1991	93	93	93	97	107	118	100
1992	93	94	95	99	102	118	100
1993	93	94	97	97	107	113	100
1994	94	91	96	95	110	114	99
1995	94	95	94	99	113	103	103
1996	97	90	96	95	113	112	98
1997	95	97	99	96	109	107	98
1998	94	93	98	96	107	115	98
1999	96	96	99	98	106	110	96
2000	93	94	101	99	110	108	94
2001	94	94	97	102	111	106	97
2002	98	95	96	99	111	107	94
2003	96	94	97	98	109	109	97
2004	94	94	99	101	110	106	96
2005	923	96	97	98	112	107	97

Appendix Table A2.4: Distance travelled by month of year (index): Great Britain, 1991-2005
Source of data: Department for Transport

Distance travelled by month of year (Index): Great Britain

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
1991	93	64	95	113	100	109	105	132	104	102	88	97
1992	80	87	98	114	107	100	109	109	111	108	94	83
1993	89	96	96	109	97	94	116	121	107	97	94	86
1994	83	85	96	103	110	102	105	124	110	92	90	99
1995	85	99	88	109	112	102	102	120	93	106	98	87
1996	81	86	91	96	106	112	109	116	111	100	97	95
1997	76	88	99	110	106	104	102	121	100	99	93	103
1998	88	92	96	108	110	97	116	118	109	93	86	90
1999	94	103	88	86	93	111	125	107	99	115	91	83
2000	78	97	102	102	101	108	105	110	105	105	106	88
2001	87	89	90	98	112	113	105	102	91	96	121	94
2002	81	99	95	108	102	102	104	117	101	103	96	85
2003	84	92	101	101	97	109	116	111	97	107	97	91
2004	91	89	97	101	102	101	104	112	100	102	102	100
2005	87	87	98	96	107	104	101	113	108	107	101	92

Appendix Table A2.5: Results of the addition of time and square of time as explanatory variable for analysing the temporal effects (Dataset 1)

Model No.	Improvement in *BIC*	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
1	184	8.61×10^{-05}	1.15	16	-1.41×10^{-08}	-10.65	16
2	290	9.90×10^{-06}	1.55	16	-1.41×10^{-08}	-12.56	16
3	1,489	7.97×10^{-06}	1.18	16	-1.40×10^{-08}	-11.78	16
4	292	9.92×10^{-06}	1.56	16	-1.42×10^{-08}	-12.56	16
5	1,410	8.23×10^{-06}	1.19	16	-1.40×10^{-08}	-11.46	16
6	2,068	9.19×10^{-06}	1.66	16	-1.41×10^{-08}	-14.43	16
7	1,925	9.45×10^{-06}	1.64	16	-1.41×10^{-08}	-13.85	16
8	2,132	9.09×10^{-06}	1.68	16	-1.41×10^{-08}	-14.67	16
9	1,964	9.46×10^{-06}	1.66	16	-1.41×10^{-08}	-14.02	16
10	2,122	9.23×10^{-06}	1.69	16	-1.41×10^{-08}	-14.62	16
11	2,127	9.23×10^{-06}	1.70	16	-1.41×10^{-08}	-14.64	16
12	191	9.23×10^{-06}	1.69	16	-1.41×10^{-08}	-14.62	16
13	214	8.68×10^{-06}	1.69	16	-1.40×10^{-08}	-15.42	16
14	224	8.71×10^{-06}	1.73	16	-1.40×10^{-08}	-15.74	16
15	234	8.66×10^{-06}	1.73	16	-1.40×10^{-08}	-15.78	16
16	205	-1.8×10^{-05}	-1.57	87	-1.35×10^{-08}	-14.80	17
17	131	-2.2×10^{-05}	-2.47	51	-1.99×10^{-08}	-11.91	56
18	190	-1.4×10^{-05}	-1.34	67	-1.33×10^{-08}	-14.26	17
19	69	1.45×10^{-05}	2.43	23	-1.71×10^{-08}	-8.81	77
20	64	-1.7×10^{-05}	-1.44	86	-2.11×10^{-08}	-8.56	125
21	59	-1.7×10^{-05}	-1.69	68	-2.17×10^{-08}	-8.26	142
22	82	-1.7×10^{-05}	-1.69	64	-2.15×10^{-08}	-9.60	103
23	84	-2×10^{-05}	-1.88	68	-1.99×10^{-08}	-9.68	87
24	53	-5.4×10^{-05}	-2.34	341	-2.48×10^{-08}	-7.90	200
25	5	-2.2×10^{-05}	-2.05	74	-1.66×10^{-08}	-3.72	402
26	2	-1.1×10^{-05}	-3.77	580	-1.49×10^{-08}	-3.31	408

Appendix Table A2.6: Results of the Park and Glejser test-Dataset 1

Tests	**Results of Test**
Park Test	u_i^2= 10.38+ 0.0035*(Estimated values of Road accidents) (3.66) (0.73) R^2= 0.01
Glejser Test	ABS(u_i)= 3.214+ 0.00047*(Estimated values of Road accidents) (8.35) (0.72) R^2= 0.01

u_i represents the deviance residuals, ( ) represents the t values.

Significant t values of the explanatory variable suggest the presence of heteroscedasticity.

Appendix Table A2.7: Results of the addition of time and square of time as explanatory variable for analysing the temporal effects (Dataset 2)

Model No.	Improvement in *BIC*	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
1	2319	-8.54×10^{-05}	-35.20	16	3.10×10^{-09}	7.20	16
2	2073	-8.49×10^{-05}	-35.92	16	2.99×10^{-09}	7.12	16
3	13835	-8.63×10^{-05}	-36.37	16	3.02×10^{-09}	7.17	16
4	2086	-8.50×10^{-05}	-35.93	16	2.99×10^{-09}	7.13	16
5	13548	-8.63×10^{-05}	-36.15	16	3.09×10^{-09}	7.29	16
6	13925	-8.59×10^{-05}	-37.17	16	2.90×10^{-09}	7.07	16
7	13644	-8.59×10^{-05}	-36.95	16	2.98×10^{-09}	7.21	16
8	13971	-8.60×10^{-05}	-37.28	16	2.92×10^{-09}	7.13	16
9	13679	-8.59×10^{-05}	-37.00	16	2.98×10^{-09}	7.23	16
10	13972	-8.59×10^{-05}	-37.22	16	2.89×10^{-09}	7.06	16
11	38	-8.59×10^{-05}	-37.22	16	2.89×10^{-09}	7.06	16
12	28	-8.63×10^{-05}	-37.61	16	2.95×10^{-09}	7.25	16
13	27	-8.62×10^{-05}	-37.62	16	2.92×10^{-09}	7.20	16
14	39	-8.62×10^{-05}	-37.63	16	2.92×10^{-09}	7.10	16
15	44	-8.47×10^{-05}	-41.68	16	2.71×10^{-09}	7.52	16
16	44	-8.7×10^{-05}	-39.86	16	2.94×10^{-09}	7.56	16
17	53	1.91×10^{-05}	9.39	17	-2.88×10^{-09}	-8.08	16
18	54	-9.3×10^{-05}	-40.65	16	3.30×10^{-09}	8.15	16
19	62	-9.4×10^{-05}	-42.17	16	3.41×10^{-09}	8.63	16
20	43	-8.8×10^{-05}	-39.39	16	2.94×10^{-09}	7.45	16
21	49	-8.8×10^{-05}	-39.08	17	3.02×10^{-09}	7.82	18
22	33	9.24×10^{-06}	4.62	17	-2.34×10^{-09}	-6.69	16
23	39	1.13×10^{-05}	5.66	17	-2.50×10^{-09}	-7.14	16
24	29	9.07×10^{-06}	4.51	17	-2.27×10^{-09}	-6.46	16
25	26	-7.4×10^{-05}	-33.65	16	2.40×10^{-09}	6.21	16
26	43	-8.6×10^{-05}	-38.18	18	2.85×10^{-09}	7.40	17
27	32	9.20×10^{-06}	4.60	17	-2.34×10^{-09}	-6.70	16
28	32	9.35×10^{-06}	4.66	17	-2.32×10^{-09}	-6.63	16
29	179	4.08×10^{-05}	19.27	19	-4.86×10^{-09}	-13.81	17
30	50	1.3×10^{-05}	6.52	17	-2.75×10^{-09}	-7.94	16
31	152	3.77×10^{-05}	17.81	19	-4.50×10^{-09}	-12.83	17
32	160	3.88×10^{-05}	18.29	19	-4.62×10^{-09}	-13.16	17
33	-6	1.71×10^{-07}	0.07	26	-8.53×10^{-10}	-2.40	19

Appendix Table A2.8: Number of road accidents occurring on each day in different bands (Dataset 2)

Range	Comparison of number of observations	
	Observed Accidents	Estimated Accidents
Under 50	273,729	273,957
50 to 100	2,785	3,562
100 to 150	2,706	1,910
150 to 200	206	-
Greater than 200	3	-
Total	279,429	279,429

Appendix Table A2.9: Number of zero observations in different police forces Dataset 2

Police forces with number of observations with zero values			
Police force	Number of observations	Police force	Number of observations
3	59	37	30
7	1	40	50
10	1	44	1
11	53	45	1
12	6	48	2173
14	2	53	43
16	3	54	25
17	95	55	17
21	1	60	20
22	2	61	135
23	30	63	71
30	3	91	664
31	3	92	195
32	15	93	272
33	3	94	789
34	33	95	6
35	9	96	807
36	9	98	1645

Appendix Table A2.10: Results of the Park and Glejser test-Dataset 2

Tests	**Results of Test –Dataset 2**
Park Test	u_i^2= -0.15+ 0.055*(Estimated values of Road accidents) (-66.26) (54.11) R^2= 0.0104
Glejser Test	ABS(u_i)= -0.01 + 0.014*(Estimated values of Road accidents) (-31.88) (4.77) R^2= 0.0001

u_i represents the deviance residuals, ( ) represents the t values.

Significant t values of the explanatory variable suggest the presence of heteroscedasticity.

Appendix Table A2.11: Results of Model 19-GEE-negative binomial with autoregressive error terms (Dataset 1)

Variable	Negative binomial regression		
	Coefficient	*S.E*	*t* value
Weekday	0.174	0.00274	63.58
Sunday	-0.146	0.00291	-50.01
Summer	*-0.024*	0.01323	*-1.84*
Autumn	*0.068*	0.04604	*1.47*
Winter	*0.039*	0.03815	*1.01*
Weekday-Summer	-0.043	0.00422	-10.27
Sunday-Summer	0.048	0.00451	10.71
Weekday-Autumn	0.023	0.00530	4.37
Sunday-Autumn	-0.028	0.00568	-4.87
Weekday-Winter	0.043	0.00465	9.22
Sunday-Winter	-0.050	0.00499	-9.93
January	*0.028*	0.03586	*0.77*
February	*-0.034*	0.03585	*-0.96*
March	0.054	0.01551	3.47
May	*0.023*	0.01549	*1.47*
July	-0.044	0.01544	-2.87
August	-0.122	0.01592	-7.68
September	0.041	0.01603	2.55
October	*-0.047*	0.05121	*-0.92*
December	*0.058*	0.03498	*1.66*
Time	-3.09×10^{-05}	0.00001	-5.75
Public Holidays	-0.216	0.01282	-16.85
Christmas Holidays	-0.426	0.03037	-14.03
New-year Holidays	-0.116	0.03303	-3.51
D.T per veh*	0.00012	0.00002	7.60
Constant	-16.461	0.26981	-61.01

Appendix Table A2.12: Results of Model 22-GEE-negative binomial with autoregressive error terms (Standard errors corrected by using White's procedure)-Dataset 2

Variable	Negative binomial regression		
	Coefficient	*S.E*	*t* value
Weekday	0.168	0.00580	28.99
Sunday	-0.140	0.00557	-25.18
Summer	*0.019*	0.04033	0.48
Autumn	0.046	0.02540	1.81
Winter	*0.018*	0.03295	0.56
Weekday-Summer	-0.050	0.00392	-12.74
Sunday-Summer	0.054	0.00388	13.80
Weekday-Autumn	0.026	0.00273	9.65
Sunday-Autumn	-0.028	0.00331	-8.33
Weekday-Winter	0.051	0.00393	12.98
Sunday-Winter	-0.058	0.00455	-12.71
January	0.064	0.03664	1.75
February	*-0.017*	0.03806	-0.45
March	0.058	0.00474	12.22
May	0.027	0.00478	5.73
June	*-0.036*	0.03523	-1.01
July	-0.080	0.03574	-2.24
August	-0.150	0.03593	-4.18
September	*0.006*	0.03610	0.17
October	*-0.030*	0.02950	-1.00
December	0.084	0.03588	2.35
Time	-4.12×10^{-6}	0.00002	-0.19
Public Holidays	-0.185	0.01689	-10.94
Christmas Holidays	-0.475	0.02566	-18.50
New-year Holidays	-0.047	0.02280	-2.08
Population density	7.58×10^{-5}	0.00002	3.36
Veh per person	-2.004	0.46739	-4.29
Constant	-13.669	0.18727	-72.99

Appendix Table 3.1: Results of models for the police forces with meteorological variables when season was carried forward from model 5 onwards instead of month (Dataset 3)

Model	**D.F**	**Scale**	**Log-Likelihood**	***BIC***	**Difference in *BIC* when month is used**
1	1	0.0630	-17,696	35,400	0
2	17	0.0396	-17,039	34,214	0
3	12	0.0642	-17,725	35,547	0
4	4	0.0660	-17,766	35,564	0
5	5	0.0489	-17,324	34,687	-52
6	6	0.0406	-17,057	34,162	-70
7	7	0.0261	-16,483	33,023	-126
8	23	0.0158	-15,826	31,836	-232
9	8	0.0258	-16,473	33,010	-107
10	8	0.0261	-16,483	33,030	-182
11	8	0.0260	-16,480	33,024	-125
12	8	0.0258	-16,473	33,010	-106
13	8	0.0261	-16,483	33,031	128
14	9	0.0254	-16,450	32,971	-127
15	9	0.0260	-16,480	33,031	-177
16	9	0.0258	-16,471	33,014	-110
17	9	0.0258	-16,473	33,018	-108
18	10	0.0254	-16,449	32,979	-127
19	10	0.0257	-16,467	33,015	-152
20	10	0.0258	-16,471	33,022	-110
21	11	0.0253	-16,448	32,984	-125
22	11	0.0257	-16,463	33,014	-175
23	12	0.0252	-16,443	32,982	-146
24	28	0.0142	-15,701	31,626	-153

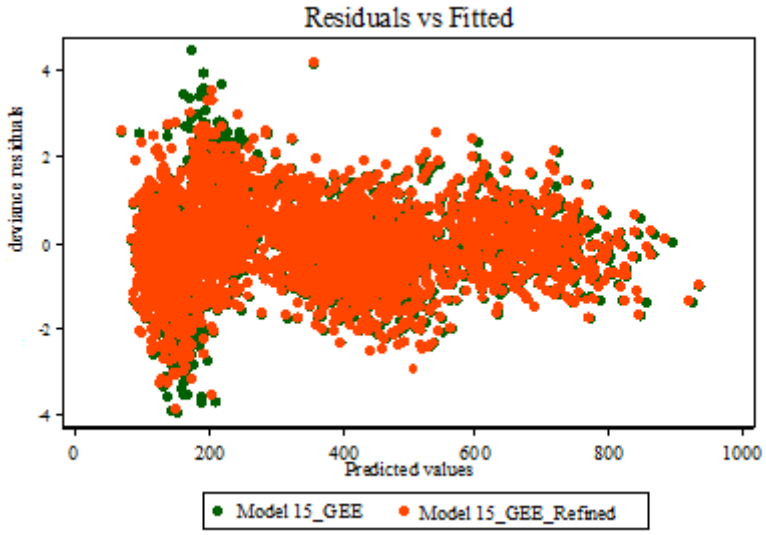
BIC represents the Bayesian information criterion

Negative sign in the difference in BIC shows that BIC is less preferable compared to same model when month was used instead of season in the linear predictor. This also suggests that month performed better than season in each model.

Appendix Table 3.2: Results of the addition of time and square of time as explanatory variable for analysing the temporal effects (Dataset 3)

Model No.	Improvement in *BIC*	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
1	840	-2.4×10^{-03}	-7.54	16	-6.0×10^{-07}	-0.35	16
2	1,529	-2.4×10^{-03}	-10.72	16	-6.3×10^{-07}	-0.52	16
3	903	-2.4×10^{-03}	-7.76	16	-8.3×10^{-07}	-0.50	16
4	869	-2.4×10^{-03}	-7.57	16	-8.5×10^{-07}	-0.50	16
5	-8	-2.4×10^{-03}	-7.76	16	-8.3×10^{-07}	-0.50	16
6	--8	-2.6×10^{-03}	-9.03	16	-2.5×10^{-07}	-0.16	16
7	-6	-1.4×10^{-03}	-5.87	16	1.6×10^{-06}	1.28	16
8	-5	-1.3×10^{-03}	-6.73	17	1.8×10^{-06}	1.84	16
9	-6	-1.4×10^{-03}	-5.96	16	1.7×10^{-06}	1.35	16
10	-5	-1.5×10^{-03}	-6.28	16	2.1×10^{-06}	1.65	16
11	-6	-1.4×10^{-03}	-6.06	16	1.8×10^{-06}	1.44	16
12	-6	-1.4×10^{-03}	-5.95	16	1.8×10^{-06}	1.39	16
13	-6	-1.4×10^{-03}	-5.94	16	1.7×10^{-06}	1.35	16
14	-6	-1.4×10^{-03}	-6.17	16	2.0×10^{-06}	1.59	16
15	-5	-1.5×10^{-03}	-6.34	17	2.1×10^{-06}	1.71	16
16	-5	-1.5×10^{-03}	-6.20	17	2.1×10^{-06}	1.63	17
17	-6	-1.4×10^{-03}	-6.02	17	1.9×10^{-06}	1.47	16
18	-5	-1.4×10^{-03}	-6.18	17	2.0×10^{-06}	1.61	16
19	-5	-1.5×10^{-03}	-6.36	17	2.2×10^{-06}	1.75	17
20	-5	-1.5×10^{-03}	-6.23	17	2.1×10^{-06}	1.67	17
21	-5	-1.5×10^{-03}	-6.29	17	2.2×10^{-06}	1.77	17
22	-6	-1.4×10^{-03}	-6.19	17	2.0×10^{-06}	1.57	17
23	-5	-1.4×10^{-03}	-6.13	17	2.0×10^{-06}	1.59	17
24	-2	-1.3×10^{-03}	-7.14	17	2.4×10^{-06}	2.49	17

Appendix Table A3.3: Comparison of results of deviance residuals of model 15 with the model 15 refinned* (GEE-AR1) -Dataset 3



Model 15 refined is the model in which two more explanatory variables of Rain in Cambridgeshire and minimum monthly temperature in Grampian are added.

Appendix Table A3.4: Results of the Park and Glejser test-Dataset 3

Test	**Results of Test**
Park Test	u_i^2= 1.82 -0.002*(Estimated values of Road accidents) (26.85) (-12.49) R^2= 0.05
Glejser Test	ABS(u_i)= 1.077 -0.0007*(Estimated values of Road accidents) (42.79) (12.65) R^2= 0.05

u_i represents the deviance residuals, ( ) represents the t values.

Significant t values of the explanatory variable suggest the presence of heteroscedasticity.

Appendix Table A3.5: Results of Model 15-GEE-negative binomial with autoregressive error terms (Standard errors corrected by using White's procedure)

Variable	Negative binomial regression		
	Coefficient	*S.E*	*t* value
January	0.061	0.0197	3.11
February	0.025	0.0214	1.17
March	-0.069	0.0116	-5.98
April	-0.122	0.0082	-14.83
May	-0.056	0.0100	-5.64
June	-0.011	0.0151	-0.75
July	-0.022	0.0281	-0.78
August	-0.081	0.0296	-2.74
September	0.031	0.0128	2.41
October	0.025	0.0078	3.27
November	0.131	0.0159	8.26
December	0.088	0.0229	3.84
Time	-0.001	0.0003	-3.02
Population density	0.0002	0.0001	3.15
Vehicle per person	-1.290	0.1111	-11.61
Mean Min Temperature	-0.008	0.0038	-2.04
Rain	0.001	0.00005	11.33
Constant	-13.908	0.0681	-204.38

Appendix Table A4.1: Results of the addition of time and square of time as explanatory variable for analysing the temporal effects (Dataset 4)

Model No.	Improvement in *BIC*	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
1	373	-3.4×10^{-05}	-0.79	16	-9.8×10^{-08}	-4.24	16
2	835	2.8×10^{-05}	0.84	16	-1.4×10^{-07}	-8.23	16
3	292	-6.7×10^{-05}	-1.92	16	-4.8×10^{-08}	-2.58	16
4	868	-7.2×10^{-05}	-3.33	16	-4.9×10^{-08}	-4.28	16
5	8	-7.2×10^{-05}	-3.33	16	-4.9×10^{-08}	-4.28	16
6	6	-7.3×10^{-05}	-3.35	16	-4.9×10^{-08}	-4.14	16
7	24	-4.5×10^{-05}	-2.17	16	-6.5×10^{-08}	-5.91	16
8	1	-1.1×10^{-04}	-5.18	16	-4.0×10^{-08}	-3.46	16
9	14	-8.5×10^{-05}	-4.17	16	-5.4×10^{-08}	-5.01	16
10	14	-9.1×10^{-05}	-4.48	16	-5.4×10^{-08}	-5.00	16
11	14	-9.1×10^{-05}	-4.50	16	-5.3×10^{-08}	-4.94	16
12	14	-9.3×10^{-05}	-4.64	16	-5.2×10^{-08}	-4.93	16
13	15	-9.0×10^{-05}	-4.51	16	-5.4×10^{-08}	-5.08	16
14	16	-8.8×10^{-05}	-4.43	16	-5.4×10^{-08}	-5.13	16
15	22	-8.8×10^{-05}	-5.43	16	-5.0×10^{-08}	-5.75	16
16	21	-8.8×10^{-05}	-5.53	16	-4.8×10^{-08}	-5.60	16
17	20	-9.2×10^{-05}	-5.76	16	-4.7×10^{-08}	-5.53	16

Appendix Table A4.2: Results of the Park and Glejser test-Dataset 4

Tests	**Results of Test**
Park Test	u_i^2= 1.27 -0.00287*(Estimated values of number of vehicles involved in road accidents) (128.57) (-28.25) R^2= 0.017
Glejser Test	ABS(u_i)= 0.900 -0.0012*(Estimated values of number of vehicles involved in road accidents) (257.77) (-34.77) R^2= 0.026

u_i represents the deviance residuals, ( ) represents the t values.

Significant t values of the explanatory variable suggest the presence of heteroscedasticity.

Appendix Table A4.3: Results of Model 17-GEE-negative binomial with autoregressive error terms (Standard errors corrected by using White's procedure)-Dataset 4

Variable	GEE-AR1 Negative binomial regression		
	Coefficient	*S.E*	*t* value
Motorway	-1.092	0.00140	-780.04
Rural A	-0.288	0.00029	-987.06
Rural Minor	-0.160	0.00040	-401.18
Urban Minor	0.725	0.00077	940.35
Pedal cycle	0.891	0.00131	679.04
Motorcycle	0.773	0.00012	6524.29
Bus	-0.042	0.00117	-35.51
Goods vehicle	-0.484	0.00014	-3394.11
Time	-0.00018	0.00004	-4.47
Weekday 1	0.155	0.03934	3.94
Weekday 2	0.105	0.04694	2.24
Sunday	-0.197	0.06592	-2.98
Summer	0.106	0.03335	3.17
Autumn	-0.095	0.01839	-5.14
Winter	*0.030*	0.04454	*0.67*
January	-0.079	0.01999	-3.95
February	-0.076	0.02233	-3.40
March	-0.063	0.02292	-2.76
May	0.053	0.01454	3.63
June	-0.051	0.01369	-3.74
July	-0.066	0.01296	-5.09
August	-0.074	0.01250	-5.89
October	0.160	0.02780	5.76
WD1-Summer	-0.047	0.01301	-3.60
WD2-Summer	-0.043	0.01196	-3.59
Sun-Summer	0.074	0.01970	3.76
WD1-Autumn	0.033	0.00843	3.90
WD2-Autumn	0.028	0.00856	3.26
Sun-Autumn	-0.039	0.01487	-2.65
WD1-Winter	0.034	0.01231	2.78
WD2-Winter	0.039	0.01275	3.02
Sun-Winter	-0.066	0.01930	-3.41
Holidays	-0.146	0.03845	-3.78
New-year	-0.300	0.06505	-4.62
Christmas	-0.276	0.05718	-4.82
MC.Mot	-0.093	0.00157	-59.63
Bus.Mot	-1.075	0.00214	-501.54
GV.Mot	0.238	0.00247	96.31
PC.RA	0.730	0.00375	195.00
MC.RA	0.248	0.00922	26.85
Bus.RA	-0.644	0.00146	-441.58
GV.RA	0.252	0.00264	95.69
PC.RM	-1.047	0.00178	-588.94
MC.RM	*0.004*	0.00903	*0.41*
Bus.RM	-0.755	0.00119	-636.96
GV.RM	0.830	0.00140	591.26
PC.UM	-0.473	0.00400	-118.13

MC.UM	0.140	0.00296	47.16
Bus.UM	0.075	0.00249	30.10
GV.RM	1.080	0.00251	430.66
MC-Rural-Sunday	0.899	0.09290	9.67
Constant	-14.270	0.13207	-108.05

Appendix A5.1

H-likelihood:

H-likelihood being fundamentally different from classical likelihood has generated some controversies. The h-likelihood is a special kind of extended likelihood, where the scale of random parameter is specified to certain conditions. For inference in HGLM three likelihoods are available, the h-likelihood and two adjusted profile likelihoods namely the marginal likelihoods, and the restricted or residual likelihoods. The marginal likelihood is an adjusted profile likelihood, eliminating nuisance random effects v from h by integration while restricted likelihood for mixed linear models is that which eliminates nuisance fixed effects β from L by conditioning on the ML estimates of β. H-likelihood is used for inference about the v, the marginal likelihood L for β and the adjusted profile likelihood for the dispersion parameters. Table 5.1 shows the likelihoods which are used to compare the fixed, random, and dispersion parts of the HGLM model.

1. H-likelihood for random part: Lee et al (2006, 188, ff) give the following expression for h-likelihood of random part in HGLM with Poisson-gamma distribution and log link;

$$h = \sum_{ij}\{y_{ij}\mathbf{x}_{\mathbf{ij}}^{\mathbf{t}}\boldsymbol{\beta} - \log\Gamma(y_{ij}+1)\} + \sum_{i}\left[(y_{i+}+\frac{1}{\lambda})v_i - (\mu_{i+}+\frac{1}{\lambda})u_i - \frac{\log(\lambda)}{\lambda} - \log\{\Gamma(\frac{1}{\lambda})\}\right]$$

Where $y_{i+} = \sum_j y_{ij}$ and $\mu_{i+} = \sum_j \exp \mathbf{x}_{\mathbf{ij}}^{\mathbf{t}}\boldsymbol{\beta}$

and they adopted Stirling approximation by $\log\Gamma(x)$:

$$\log\Gamma(x) = (x-\frac{1}{2})\log(x) + \frac{\log(2\pi)}{2} - x$$

In these equations

i represents the index of the random part in the model,

v_i represents the random coefficient produced by the model,

u_i represents the exponential of v_i coefficient of random part, and

λ represents the variation between random parts.

2. H-likelihood for fixed part: Lee et al., (2006, 188, ff) give the following expression for h-likelihood of the fixed part in HGLM with Poisson-gamma distribution and log link is estimated as below:

$$p_v(h) = \sum_{ij}\{y_{ij}\mathbf{x_{ij}^t}\boldsymbol{\beta} - \log\Gamma(y_{ij}+1)\} + \sum_i[-(y_{i+}+\frac{1}{\lambda})\log(\mu_{i+}+\frac{1}{\lambda}) + (y_{i+}+\frac{1}{\lambda})\log(y_{i+}+\frac{1}{\lambda}) - (y_{i+}+\frac{1}{\lambda}) - \frac{\log(\lambda)}{\lambda} - \log\{\Gamma(\frac{1}{\lambda})\} - \frac{1}{2}\log(y_{i+}+\frac{1}{\lambda}) + \frac{1}{2}\log(2\pi)]$$

3. H-likelihood for dispersion part

H-likelihood of dispersion part in HGLM with Poisson-gamma distribution and log link is estimated as under (Lee et al., 2006, 188, ff);

$$p_{\beta,v}(h) = (h - [\log\det\{D(h,(\beta,v))/2\pi\}]/2\Big|_{\beta=\hat{\beta},v=\hat{v}} \qquad 5\text{-}16$$

where

$$D(h,(\beta,v)) = T_M{}^t\sum\nolimits_M^{-1}T_M$$

$$D = -\begin{bmatrix} \frac{\partial^2 h}{\partial v_i \partial v_j} & \frac{\partial^2 h}{\partial \beta_k \partial v_i} \\ \frac{\partial^2 h}{\partial v_i \partial \beta_k} & \frac{\partial^2 h}{\partial \beta_k \partial \beta_l} \end{bmatrix} \qquad -\frac{\partial^2 h}{\partial v_i \partial v_j} = \begin{bmatrix} (y_{i+}+1/\lambda) & 0 \\ 0 & (y_{j+}+1/\lambda) \end{bmatrix}$$

$$-\frac{\partial^2 h}{\partial v_i \partial \beta_l} = \sum_j x_{lij}\mu_{ij} \qquad -\frac{\partial^2 h}{\partial \beta_k \partial \beta_l} = \sum_{ij} x_{lij}x_{kij}\mu_{ij}$$

4. Extended quasi likelihood (EQL):

Nelder and Pregibon (1987) introduced quasi likelihood function which allows for the comparison of various forms of all components of a generalized linear model, i.e. the linear predictor, the link function, and the variance function. The contribution of y_i to the EQL (Lee et al., 2006, 85, ff) is:

$$Q_i\left(\mu_i, \phi; y_i\right) = -\frac{1}{2}\log\left(\phi V\left(y_i\right)\right) - \frac{1}{2\phi} d\left(y_i, \mu_i\right)$$

and the total is denoted by $q^+ = \sum_i Q_i$, where $d\left(y_i, \mu_i\right)$ is the deviance function defined by

$$d_i \equiv d\left(y_i, \mu_i\right) = 2\int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)}\, du$$

EQL forms the basis for the joint modelling of structured mean and dispersion parameters, both within the GLM framework. EQL for the Poisson-gamma, when the $y \mid u$ is Poisson with mean u, and u itself is gamma with density

$$f\left(u\right) = \frac{1}{\Gamma(v)} \frac{1}{\alpha}\left(\frac{u}{\alpha}\right)^{v-1} \exp\left(\frac{-u}{\alpha}\right), \quad u > 0$$

So that $E(y) = u = \alpha\, v$

$$\operatorname{var}(y) = \alpha v + \alpha^2 v = \mu(1+\alpha)$$

The EQL is

$$q^+(\mu, \alpha) = -\frac{n}{2}\log\left(1+\alpha\right) - \frac{1}{1+\alpha}\sum_i \left(y_i \log y_i - y_i \log \mu - y_i + \mu\right)$$

Appendix Table A5.2: h-likelihood results of Full model with different offsets (Dataset 5: Car)

Variables used				
Fixed Part	Age group, gender, age group.gender, Day of week , Month, Time ,Public holidays, New-year holidays, Christmas holidays			
Random Part	Month.Year			
Dispersion Part	Age group, gender, Day of week, Month			
H-likelihood with different offsets				
Model	Offset	Fixed	Random	Dispersion
A1:	Population	201,681	201,257	201,708
A2:	Total daily distance travelled	202,778	202,357	202,784
A3:	Yearly distance travelled per person	202,659	202,238	202,667

Appendix Table A5.3: Results of the addition of square of time as explanatory variable for analysing the temporal effects-HGLM Full model (Dataset 5)

Model No.	Improvement in h-likelihood for fixed part	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
Full Model	0	-0.0001122	-3.12	16.05	$8.16x10^{-09}$	0.45	16.08

Appendix Table A5.4: Results of the addition of square of time as explanatory variable for analysing the temporal effects-HGLM Full models (Dataset 6-9)

Model No.	Improvement in h-likelihood for fixed part	Time			Square of Time		
		Coefficient	*t* value	VIF	Coefficient	*t* value	VIF
Walk	1	-0.0001	-3.12	16	$8.16x10\text{-}^{09}$	0.45	16
Bicycle	7	-0.0003	-4.79	16	$1.32x10^{-07}$	3.51	16
Motorcycle	4	0.00003	0.45	16	$-8.48x10^{-08}$	-2.38	16
Bus	1	-0.00006	-0.82	16	$-4.30x10^{-08}$	-0.98	16

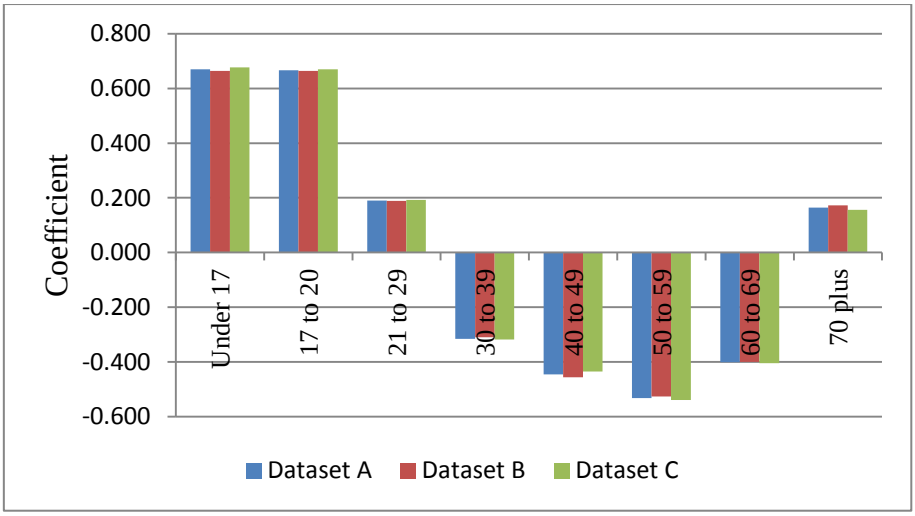
Appendix Table A5.5: Comparison of coefficients and *t* values of full model HGLM

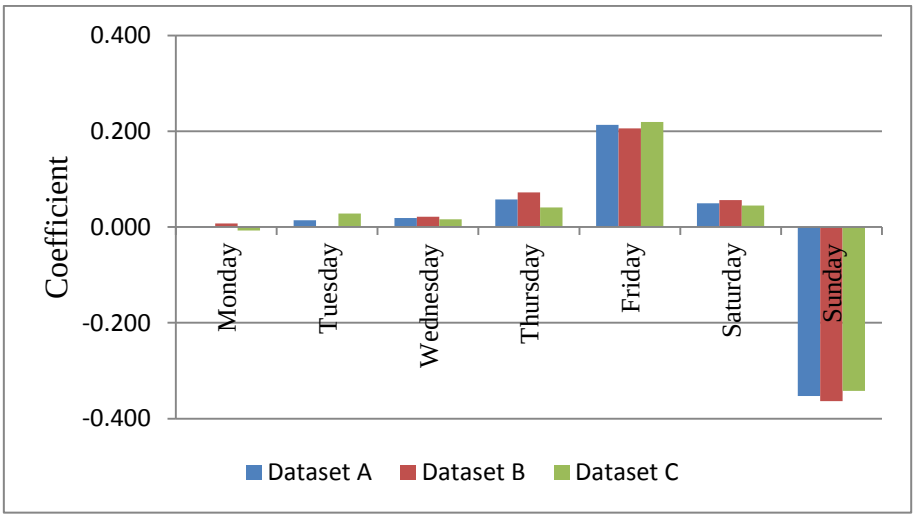
(Split sample Data: Walk)

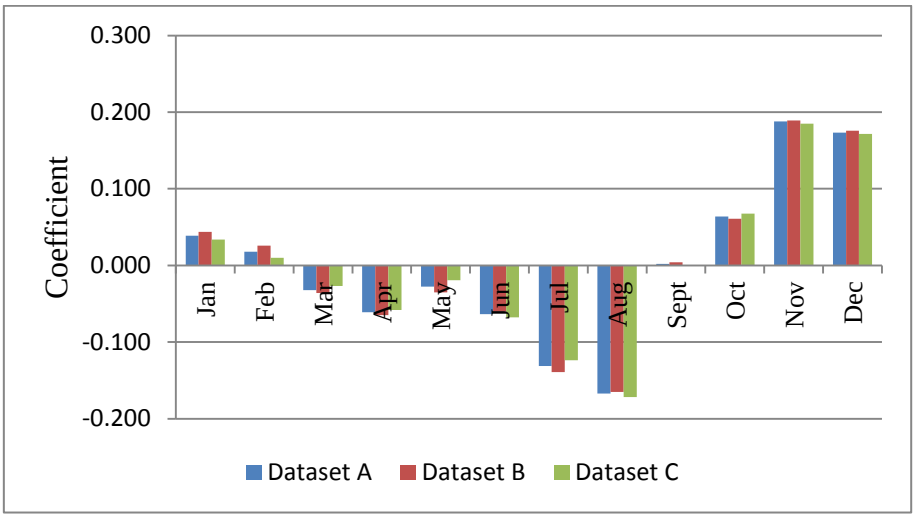
Variables	Comparison of the coefficients and t values of the Models						
	Model A		Model B		Model C		
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	t_{BC}
Under 17	0.670	65.35	0.677	46.35	0.664	45.93	0.639
17-29	0.667	50.59	0.670	35.88	0.664	35.72	0.209
30-39	-0.315	-26.55	-0.318	-19.12	-0.309	-18.27	-0.350
40-49	-0.445	-34.28	-0.434	-24.02	-0.455	-24.46	0.801
50-59	-0.531	-38.50	-0.538	-27.38	-0.526	-27.16	-0.445
60-69	-0.401	-26.04	-0.404	-18.10	-0.399	-18.81	-0.163
70 plus	0.165	14.20	0.156	9.59	0.173	10.45	-0.716
Gender	0.176	53.97	0.173	37.32	0.179	38.96	-1.002
Under 17.Male	-0.052	-3.76	-0.057	-2.91	-0.046	-2.38	-0.383
17-20. Male	*-0.005*	*-0.27*	*-0.002*	*-0.08*	*-0.010*	*-0.40*	0.225
30-39. Male	0.223	14.49	0.218	10.05	0.224	10.24	-0.196
40-49. Male	0.066	3.80	0.075	3.12	0.061	2.48	0.420
50-59. Male	-0.078	-4.15	-0.100	-3.64	-0.057	-2.19	-1.123
60-69. Male	-0.169	-7.76	-0.148	-4.79	-0.192	-6.23	1.006
70 plus. Male	-0.063	-3.67	-0.076	-3.14	-0.047	-1.96	-0.857
Monday	*0.000*	*-0.02*	*-0.008*	*-0.68*	*0.008*	*0.70*	-0.977
Tuesday	*0.014*	*1.91*	0.028	2.68	*0.000*	*-0.01*	1.879
Wednesday	0.019	2.50	*0.016*	*1.49*	0.021	2.05	-0.355
Thursday	0.058	7.88	0.041	3.94	0.072	6.99	-2.138
Saturday	0.049	6.43	0.045	4.17	0.056	5.10	-0.714
Sunday	-0.353	-36.62	-0.342	-25.59	-0.364	-26.17	1.114
January	0.039	2.32	*0.034*	*1.60*	0.044	2.37	-0.353
February	*0.018*	*1.08*	*0.010*	*0.48*	*0.026*	*1.46*	-0.590
March	*-0.032*	*-1.95*	*-0.027*	*-1.29*	-0.036	-2.01	0.323
April	-0.061	-3.62	-0.058	-2.74	-0.065	-3.40	0.232
May	*-0.028*	*-1.66*	*-0.019*	*-0.90*	*-0.035*	*-1.90*	0.559
June	-0.064	-3.81	-0.068	-3.25	-0.060	-3.19	-0.268
July	-0.131	-7.73	-0.124	-5.71	-0.139	-7.39	0.526
August	-0.167	-9.89	-0.172	-8.01	-0.165	-8.79	-0.228
September	*0.002*	*0.12*	*0.000*	*-0.01*	*0.004*	*0.23*	-0.160
October	0.064	3.89	0.068	3.24	0.061	3.43	0.237
December	0.173	10.21	0.172	7.99	0.176	9.27	-0.146
Time	0.000	-12.79	0.000	-10.70	0.000	-10.96	-0.801
Holidays	-0.135	-10.55	-0.124	-7.05	-0.148	-7.95	0.910
New_Year	*0.057*	*1.79*	*0.104*	*2.38*	*0.010*	*0.22*	1.455
Christmas	-0.356	-8.96	-0.335	-5.95	-0.370	-6.65	0.450
Constant	-13.79	-274.47	-13.75	-195.44	-13.87	-194.28	1.661

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.1: Comparison of coefficients of full model HGLM for coefficient validation: Walk





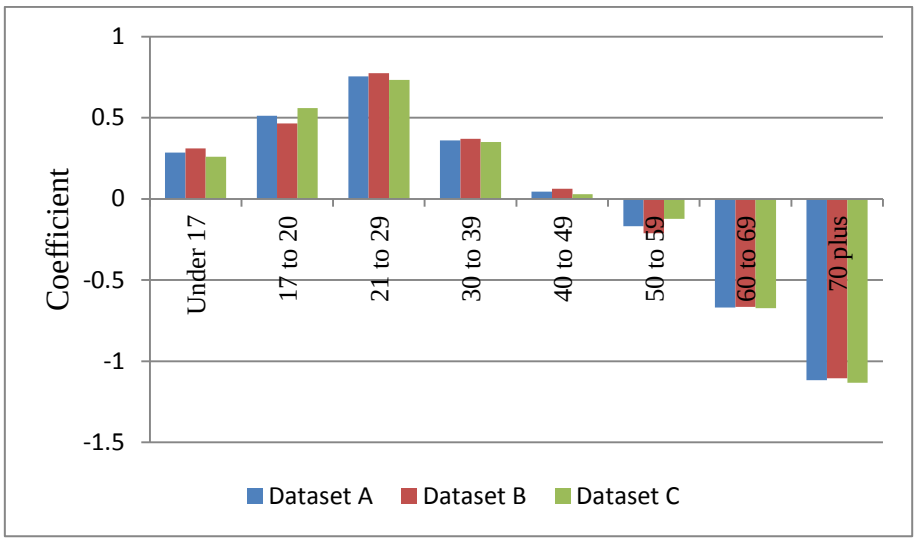


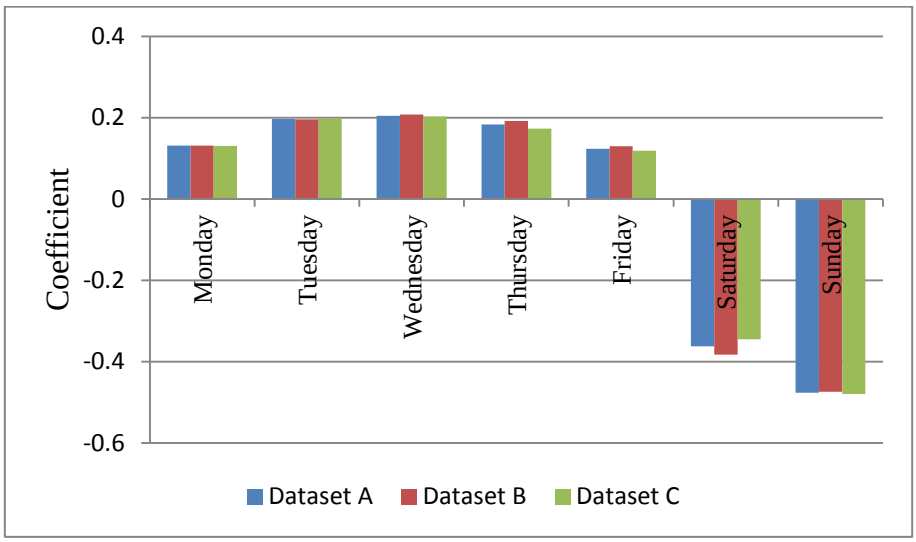
Appendix Table A5.6: Comparison of coefficients and t values of full model HGLM (Split sample Data: Bicycle)

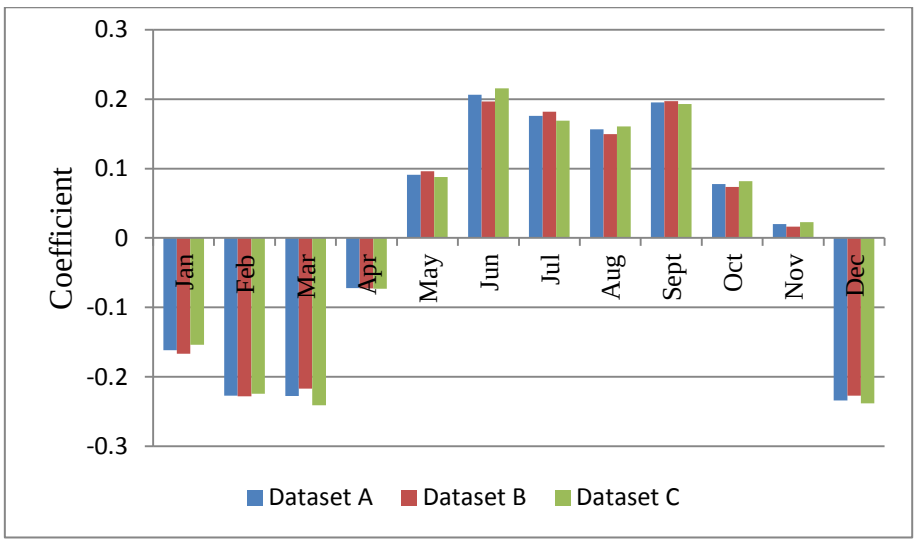
Variables	Comparison of the coefficients and t values of the Models						
	Model A		Model B		Model C		
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	t_{BC}
Under 17	0.286	13.28	0.311	9.97	0.260	8.69	1.171
17-29	0.512	18.89	0.465	12.07	0.559	14.67	-1.732
30-39	0.359	18.39	0.370	13.34	0.351	12.73	0.481
40-49	*0.044*	*1.93*	*0.062*	*1.9*	*0.029*	*0.90*	0.723
50-59	-0.168	-6.67	-0.212	-5.87	-0.124	-3.52	-1.741
60-69	-0.670	-19.89	-0.666	-13.62	-0.673	-14.48	0.117
70 plus	-1.117	-29.03	-1.106	-20.15	-1.133	-21.04	0.358
Gender	0.709	122.02	0.724	87.55	0.694	85.00	2.593
Under 17.Male	0.213	8.85	0.184	5.31	0.241	7.15	-1.176
17-20. Male	0.109	3.58	0.153	3.53	*0.063*	*1.47*	1.468
30-39. Male	*0.025*	*1.12*	*0.010*	*0.3*	*0.039*	*1.24*	-0.665
40-49. Male	*0.013*	*0.48*	*-0.012*	*-0.33*	*0.035*	*0.93*	-0.891
50-59. Male	-0.223	-7.55	-0.173	-4.09	-0.276	-6.64	1.742
60-69. Male	-0.192	-4.83	-0.201	-3.51	-0.182	-3.29	-0.234
70 plus. Male	0.358	8.14	0.375	5.99	0.349	5.66	0.294
Monday	0.131	12.87	0.131	9.22	0.131	8.98	0.029
Tuesday	0.197	20.5	0.196	14.58	0.199	14.42	-0.135
Wednesday	0.204	21.16	0.208	15.08	0.203	14.93	0.248
Thursday	0.183	18.73	0.192	14.42	0.173	12.02	0.964
Saturday	-0.362	-29.5	-0.383	-21.92	-0.345	-19.90	-1.532
Sunday	-0.476	-36.68	-0.474	-26.08	-0.479	-25.73	0.204
January	-0.162	-4.52	-0.167	-4.46	-0.154	-3.95	-0.239
February	-0.227	-6.31	-0.228	-5.97	-0.224	-5.72	-0.068
March	-0.228	-6.43	-0.217	-5.89	-0.241	-6.30	0.451
April	-0.072	-2.05	-0.073	-1.97	*-0.073*	*-1.93*	0.008
May	0.091	2.61	0.096	2.65	0.088	2.36	0.161
June	0.206	5.92	0.197	5.44	0.216	5.86	-0.369
July	0.176	5.06	0.182	5.1	0.169	4.55	0.256
August	0.157	4.51	0.150	4.16	0.161	4.36	-0.214
September	0.195	5.63	0.197	5.51	0.193	5.25	0.088
October	0.078	2.23	0.074	2.05	0.082	2.21	-0.163
December	-0.234	-6.4	-0.227	-5.78	-0.238	-5.95	0.200
Time	0.000	-4.89	0.000	-4.39	0.000	-4.67	0.283
Holidays	-0.235	-12.86	-0.278	-10.37	-0.193	-7.74	-2.314
New_Year	-0.582	-6.55	-0.567	-4.89	-0.578	-4.19	0.061
Christmas	-0.523	-6.53	-0.406	-3.75	-0.646	-5.44	1.493
Constant	-15.73	-131.15	-15.66	-99.2	-15.80	-87.04	0.593

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.2: Comparison of coefficients of full model HGLM for coefficient validation: Bicycle





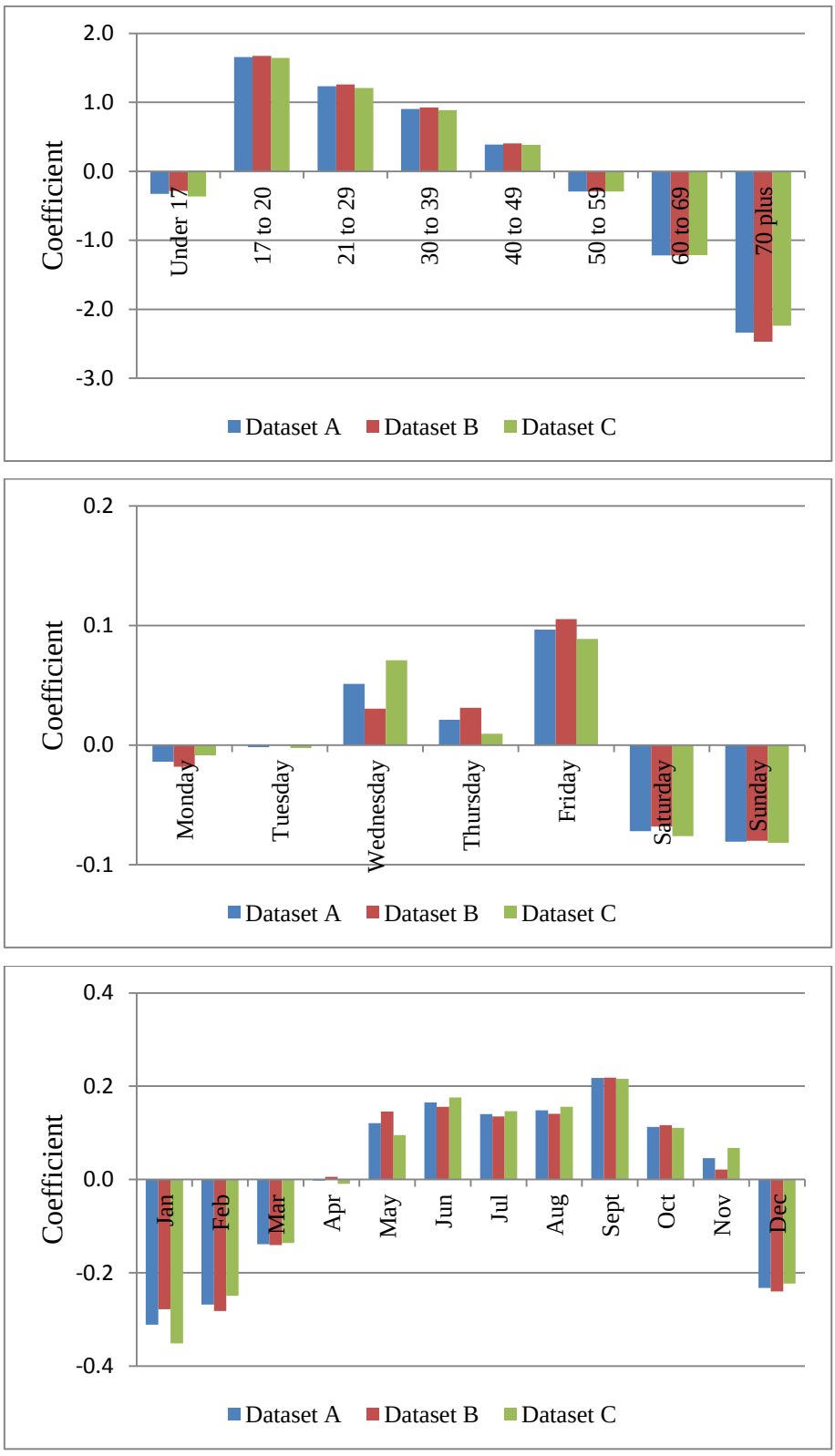


Appendix Table A5.7: Comparison of coefficients and t values of full model HGLM (Split sample Data: Motorcycle)

Variables	Comparison of the coefficients and t values of the Models						
	Model A		Model B		Model C		
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	t_{BC}
Under 17	-0.328	-12.26	-0.284	-7.59	-0.366	-9.54	1.522
17-29	1.655	74.96	1.674	53.24	1.642	52.76	0.735
30-39	0.903	44.06	0.926	31.67	0.885	30.64	1.010
40-49	0.389	15.77	0.405	11.66	0.381	10.85	0.484
50-59	-0.294	-9.58	-0.293	-6.7	-0.292	-6.74	-0.007
60-69	-1.218	-26.76	-1.218	-18.41	-1.217	-19.39	-0.008
70 plus	-2.338	-43.51	-2.471	-30.19	-2.240	-31.30	-2.124
Gender	1.036	150.91	1.045	105.04	1.031	107.91	1.030
Under 17.Male	0.000	-0.01	-0.048	-1.18	0.043	1.03	-1.563
17-20. Male	0.042	1.75	0.025	0.72	0.056	1.65	-0.651
30-39. Male	0.085	3.78	0.072	2.25	0.095	2.98	-0.497
40-49. Male	0.151	5.61	0.127	3.37	0.166	4.34	-0.712
50-59. Male	0.047	1.41	0.048	1	0.046	0.97	0.022
60-69. Male	-0.016	-0.31	-0.021	-0.29	-0.013	-0.19	-0.083
70 plus. Male	-0.127	-2.06	0.020	0.22	-0.244	-2.91	2.125
Monday	-0.014	-1.6	-0.018	-1.46	-0.009	-0.70	-0.538
Tuesday	-0.002	-0.22	0.000	-0.04	-0.002	-0.20	0.118
Wednesday	0.051	6.15	0.031	2.55	0.071	6.10	-2.421
Thursday	0.021	2.47	0.031	2.59	0.010	0.78	1.267
Saturday	-0.072	-8.07	-0.068	-5.42	-0.076	-5.99	0.442
Sunday	-0.081	-8.18	-0.080	-5.66	-0.082	-5.91	0.086
January	-0.311	-10.1	-0.278	-9.03	-0.351	-9.80	1.550
February	-0.268	-8.71	-0.282	-9.12	-0.249	-6.98	-0.693
March	-0.139	-4.57	-0.141	-4.67	-0.136	-3.87	-0.099
April	-0.002	-0.07	0.006	0.19	-0.009	-0.27	0.324
May	0.121	4.05	0.146	5	0.095	2.79	1.129
June	0.165	5.51	0.156	5.28	0.176	5.11	-0.434
July	0.141	4.71	0.135	4.62	0.146	4.31	-0.252
August	0.149	5.01	0.141	4.87	0.156	4.61	-0.335
September	0.218	7.34	0.219	7.46	0.216	6.44	0.063
October	0.113	3.79	0.116	3.99	0.111	3.28	0.116
December	-0.232	-7.49	-0.240	-7.66	-0.223	-6.16	-0.355
Time	0.000	-7.08	0.000	-6.81	0.000	-6.26	0.277
Holidays	0.030	2.46	0.032	1.92	0.024	1.32	0.353
New_Year	-0.566	-9.47	-0.602	-5.88	-0.527	-7.26	-0.603
Christmas	-0.779	-12.25	-0.704	-7.95	-0.836	-9.17	1.038
Constant	-15.68	-177.42	-15.61	-115.47	-15.71	-134.14	0.323

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.3: Comparison of coefficients of full model HGLM for coefficient validation: Motorcycle

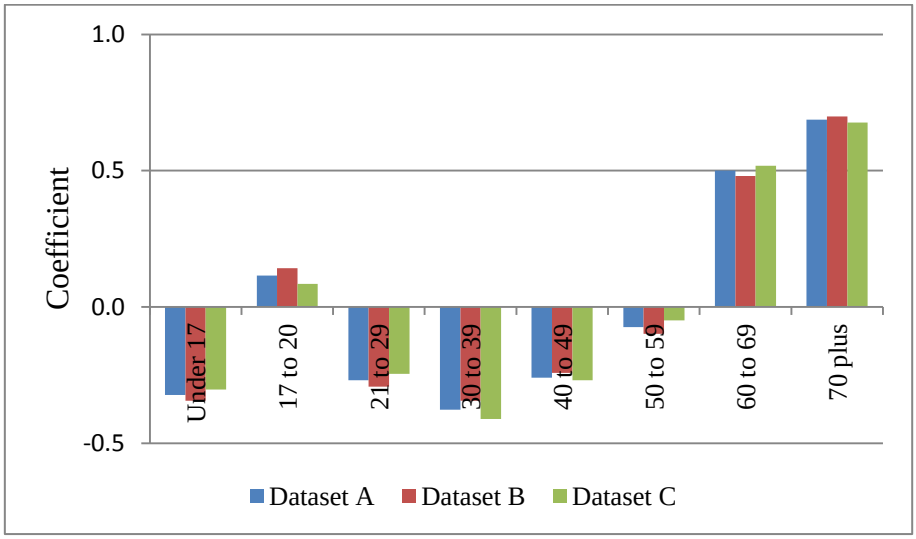


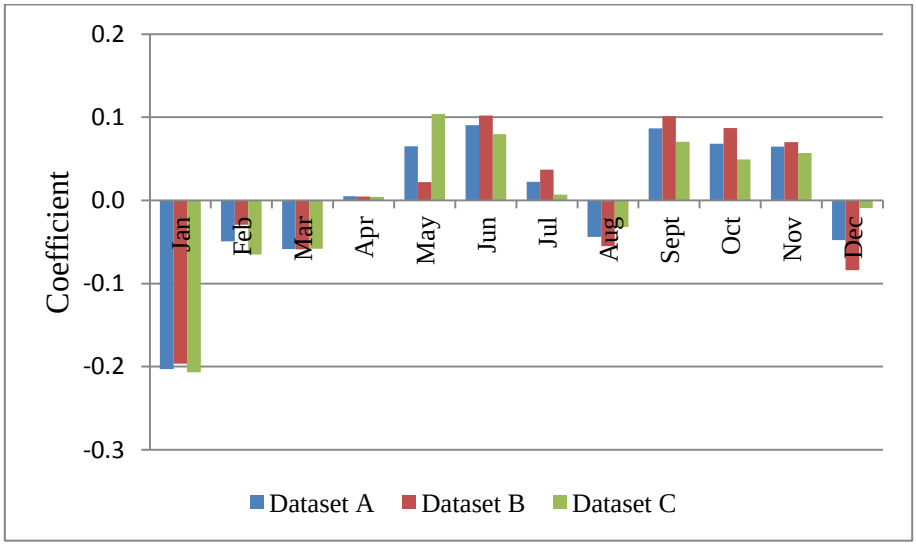
Appendix Table A5.8: Comparison of coefficients and t values of full model HGLM (Split sample Data): Bus

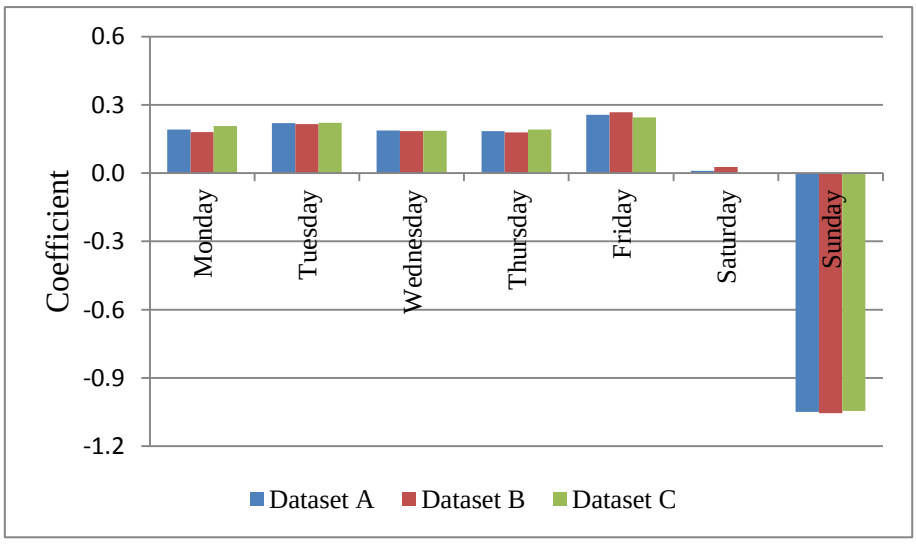
Variables	Comparison of the coefficients and t values of the Models						
	Model A		Model B		Model C		
	Coefficient	t_A	Coefficient	t_B	Coefficient	t_C	t_{BC}
Under 17	-0.323	-12.62	-0.344	-9.77	-0.303	-8.15	-0.795
17-29	0.115	4.00	0.142	3.57	0.084	2.05	1.014
30-39	-0.377	-16.86	-0.344	-11.07	-0.411	-12.79	1.499
40-49	-0.259	-12.01	-0.242	-7.93	-0.269	-8.83	0.607
50-59	-0.074	-3.60	-0.099	-3.32	-0.049	-1.72	-1.205
60-69	0.500	26.81	0.480	18.38	0.518	19.42	-1.007
70 plus	0.687	44.94	0.699	32.41	0.677	31.22	0.732
Gender	-0.490	-36.52	-0.479	-25.52	-0.501	-26.11	0.837
Under 17.Male	0.208	5.52	0.210	4.03	0.208	3.8	0.023
17-20. Male	-0.199	-4.24	-0.215	-3.32	-0.186	-2.74	-0.309
30-39. Male	0.417	13.02	0.377	8.39	0.455	9.93	-1.219
40-49. Male	0.226	7.01	0.199	4.3	0.253	5.6	-0.833
50-59. Male	-0.026	-0.79	0.026	0.56	-0.075	-1.64	1.546
60-69. Male	-0.483	-14.15	-0.468	-9.95	-0.491	-9.95	0.336
70 plus. Male	-0.331	-11.55	-0.324	-8.15	-0.342	-8.3	0.311
Monday	0.192	13.26	0.180	8.9	0.207	10.05	-0.930
Tuesday	0.219	15.16	0.216	10.62	0.221	10.72	-0.173
Wednesday	0.187	12.77	0.185	9.04	0.186	8.88	-0.017
Thursday	0.184	12.91	0.179	8.69	0.191	9.63	-0.416
Saturday	0.010	0.68	0.027	1.22	-0.003	-0.16	0.986
Sunday	-1.048	-47.75	-1.054	-33.95	-1.045	-33.7	-0.207
January	-0.203	-5.48	-0.196	-4.47	-0.207	-4.8	0.168
February	-0.049	-1.35	-0.029	-0.68	-0.065	-1.58	0.596
March	-0.058	-1.62	-0.059	-1.39	-0.058	-1.42	-0.014
April	0.005	0.13	0.005	0.1	0.004	0.1	0.007
May	0.065	1.82	0.022	0.51	0.104	2.6	-1.401
June	0.090	2.52	0.102	2.39	0.080	1.99	0.384
July	0.022	0.63	0.037	0.87	0.007	0.17	0.519
August	-0.044	-1.23	-0.055	-1.29	-0.032	-0.81	-0.396
September	0.087	2.43	0.101	2.4	0.070	1.75	0.530
October	0.068	1.90	0.087	2.08	0.049	1.2	0.649
December	-0.048	-1.31	-0.084	-1.93	-0.009	-0.22	-1.243
Time	-0.000	-6.91	-0.000	-6.37	-0.000	-5.48	-0.819
Holidays	-0.384	-12.87	-0.384	-9.26	-0.383	-8.92	-0.030
New Year	-0.446	-3.50	-0.512	-2.82	-0.377	-2.11	-0.532
Christmas	-1.167	-6.25	-1.209	-4.49	-1.144	-4.39	-0.173
Constant	-16.51	-73.37	-16.61	-51.44	-16.42	-52.33	-0.422

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.4: Comparison of coefficients of full model HGLM for coefficient validation: Bus





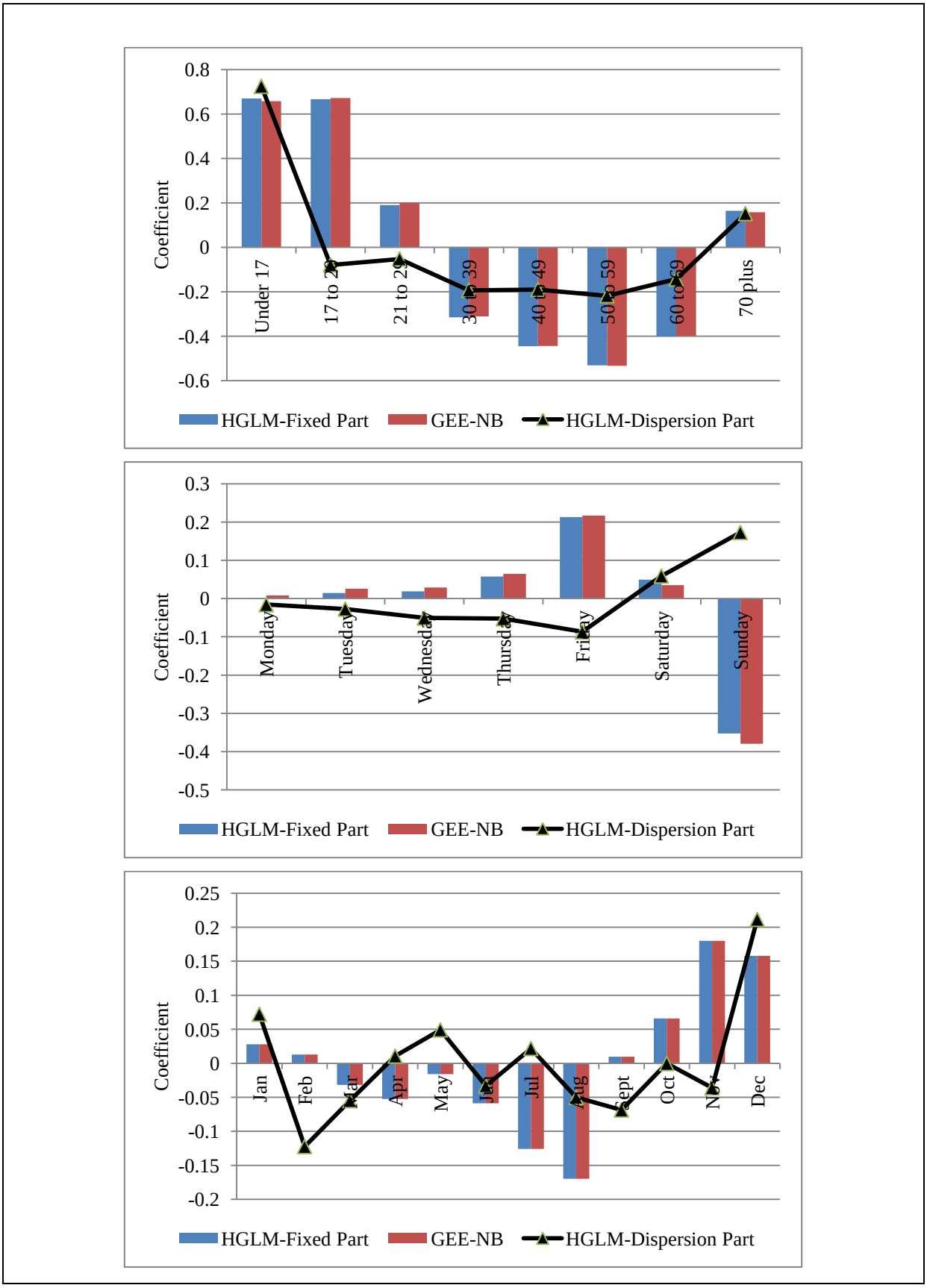


Appendix Table A5.9: Comparison of the coefficients and *t* values of some variables by HGLM and GEE-AR1 (Walk: Dataset 6)

Variable (Walk)	**Coefficient HGLM**	***t* value HGLM**	**Coefficient GEE**	***t* value GEE**
Under 17	0.670	65.35	0.657	62.13
17-20	0.667	50.59	0.672	44.71
30-39	-0.315	-26.55	-0.311	-21.59
40-49	-0.445	-34.28	-0.444	-28.76
50-59	-0.531	-38.50	-0.533	-32.48
60-69	-0.401	-26.04	-0.399	-22.89
70 plus	0.165	14.20	0.158	12.33
Gender	0.176	53.97	0.182	48.96
Under 17.Male	-0.052	-3.76	-0.056	-3.9
17-20. Male	*-0.005*	*-0.27*	*-0.001*	*-0.03*
30-39. Male	0.223	14.49	0.225	11.99
40-49. Male	0.066	3.80	0.066	3.25
50-59. Male	-0.078	-4.15	-0.080	-3.62
60-69. Male	-0.169	-7.76	-0.177	-7.37
70 plus. Male	-0.063	-3.67	-0.067	-3.7
Monday	*0.000*	*-0.02*	*0.008*	*1.07*
Tuesday	*0.014*	*1.91*	0.026	3.47
Wednesday	0.019	2.50	0.029	3.89
Thursday	0.058	7.88	0.064	8.71
Saturday	0.049	6.43	0.035	4.76
Sunday	-0.353	-36.62	-0.379	-45.18
January	0.039	2.32	0.028	2.46
February	*0.018*	*1.08*	*0.013*	*1.09*
March	*-0.032*	*-1.95*	-0.032	-2.78
April	-0.061	-3.62	-0.053	-4.48
May	*-0.028*	*-1.66*	*-0.016*	*-1.38*
June	-0.064	-3.81	-0.059	-5.02
July	-0.131	-7.73	-0.126	-10.68
August	-0.167	-9.89	-0.170	-14.18
September	*0.002*	*0.12*	*0.010*	*0.83*
October	0.064	3.89	0.066	5.89
December	0.173	10.21	0.158	14.07
Time	0.000	-12.79	0.000	-19.02
Holidays	-0.135	-10.55	-0.142	-11.53
New Year	*0.057*	*1.79*	0.073	2.41
Christmas	-0.356	-8.96	-0.348	-10.26
Constant	-13.79	-274.47	-13.77	-310.6

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.5: Comparison of coefficients by HGLM and GEE-AR1(Walk)

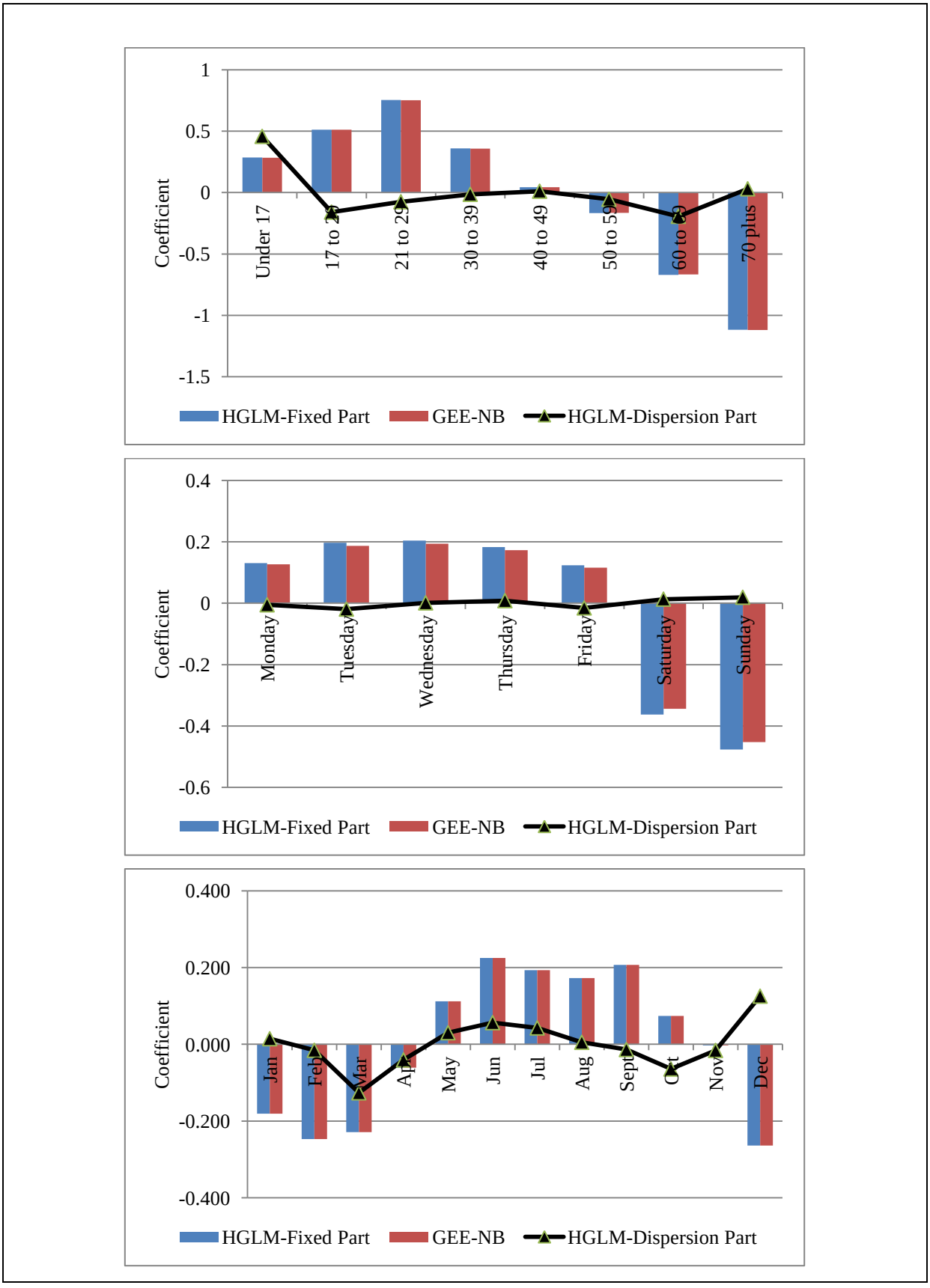


Appendix Table A5.10: Comparison of the coefficients and *t* values of the some variables by HGLM and GEE-AR1 (Bicyclists: Dataset 7)

Variable (Bicycle)	**Coefficient HGLM**	***t* value HGLM**	**Coefficient GEE**	***t* value GEE**
Under 17	0.286	13.28	0.283	13.73
17-20	0.512	18.89	0.512	16.27
30-39	0.359	18.39	0.358	16.25
40-49	*0.044*	*1.93*	*0.044*	*1.74*
50-59	-0.168	-6.67	-0.165	-5.84
60-69	-0.670	-19.89	-0.666	-16.89
70 plus	-1.117	-29.03	-1.119	-27.48
Gender	0.709	122.02	0.709	110.76
Under 17.Male	0.213	8.85	0.215	9.09
17-20. Male	0.109	3.58	0.106	2.99
30-39. Male	*0.025*	*1.12*	*0.024*	*0.93*
40-49. Male	*0.013*	*0.48*	*0.016*	*0.56*
50-59. Male	-0.223	-7.55	-0.223	-6.82
60-69. Male	-0.192	-4.83	-0.192	-4.22
70 plus. Male	0.358	8.14	0.361	7.87
Monday	0.131	12.87	0.126	12.62
Tuesday	0.197	20.5	0.187	19.39
Wednesday	0.204	21.16	0.194	20.1
Thursday	0.183	18.73	0.173	17.83
Saturday	-0.362	-29.5	-0.344	-30.4
Sunday	-0.476	-36.68	-0.452	-38.42
January	-0.162	-4.52	-0.181	-10.88
February	-0.227	-6.31	-0.247	-14.15
March	-0.228	-6.43	-0.229	-13.72
April	-0.072	-2.05	-0.061	-3.78
May	0.091	2.61	0.112	7.5
June	0.206	5.92	0.225	15.46
July	0.176	5.06	0.193	13.37
August	0.157	4.51	0.172	11.78
September	0.195	5.63	0.207	14.09
October	0.078	2.23	0.074	4.91
December	-0.234	-6.4	-0.264	-15.13
Time	0.000	-4.89	0.000	-10.39
Holidays	-0.235	-12.86	-0.222	-12.89
New_Year	-0.582	-6.55	-0.497	-7.00
Christmas	-0.523	-6.53	-0.473	-7.25
Constant	-15.73	-131.15	-15.590	-163.67

Italics indicate the non-significant t values at the 5 percent level

Appendix Figure A5.6: Comparison of coefficients by HGLM and GEE-AR1 (Bicyclists)

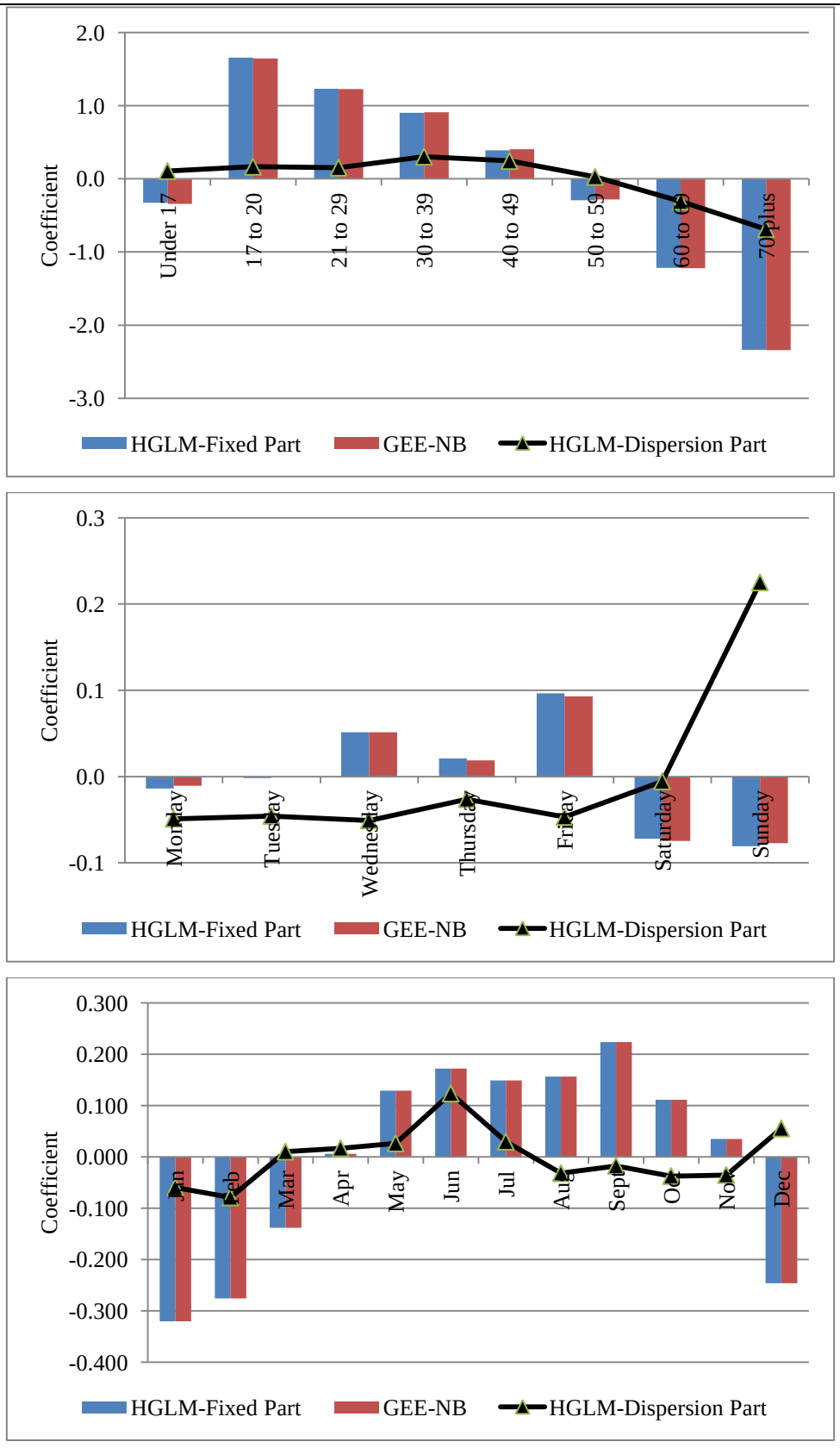


Appendix Table A5.11: Comparison of coefficient and t values of the some variables by HGLM and GEE-AR1 (Motorcyclists: Dataset 8)

Variable (MC)	**Coefficient HGLM**	***t* value HGLM**	**Coefficient GEE**	***t* value GEE**
Under 17	-0.328	-12.26	-0.341	-11.66
17-20	1.655	74.96	1.646	67.11
30-39	0.903	44.06	0.908	40.59
40-49	0.389	15.77	0.404	15.57
50-59	-0.294	-9.58	-0.282	-8.38
60-69	-1.218	-26.76	-1.220	-21.49
70 plus	-2.338	-43.51	-2.341	-29.13
Gender	1.036	150.91	1.035	120.43
Under 17.Male	*0.000*	*-0.01*	*0.011*	*0.34*
17-20. Male	*0.042*	*1.75*	*0.047*	*1.77*
30-39. Male	0.085	3.78	0.074	3
40-49. Male	0.151	5.61	0.136	4.83
50-59. Male	*0.047*	*1.41*	*0.042*	*1.17*
60-69. Male	*-0.016*	*-0.31*	*-0.008*	*-0.13*
70 plus. Male	-0.127	-2.06	*-0.121*	*-1.37*
Monday	*-0.014*	*-1.6*	*-0.011*	*-1.24*
Tuesday	*-0.002*	*-0.22*	*0.000*	*-0.03*
Wednesday	0.051	6.15	0.051	6.06
Thursday	0.021	2.47	0.019	2.21
Saturday	-0.072	-8.07	-0.075	-8.54
Sunday	-0.081	-8.18	-0.077	-8.86
January	-0.311	-10.1	-0.321	-23.71
February	-0.268	-8.71	-0.276	-20.16
March	-0.139	-4.57	-0.138	-11.02
April	*-0.002*	*-0.07*	*0.006*	*0.47*
May	0.121	4.05	0.129	11.07
June	0.165	5.51	0.172	14.78
July	0.141	4.71	0.149	12.86
August	0.149	5.01	0.156	13.53
September	0.218	7.34	0.224	19.4
October	0.113	3.79	0.111	9.49
December	-0.232	-7.49	-0.246	-18.28
Time	0.000	-7.08	0.000	-15.06
Holidays	0.030	2.46	*0.023*	*1.9*
New Year	-0.566	-9.47	-0.517	-9.97
Christmas	-0.779	-12.25	-0.750	-14.53
Constant	-15.68	-177.42	-15.63	-214.6

Italics indicate the non-significant t values with 5 percent level

Appendix Figure A5.7: Comparison of coefficients by HGLM and GEE-AR1 (Motorcyclists: Dataset 8)

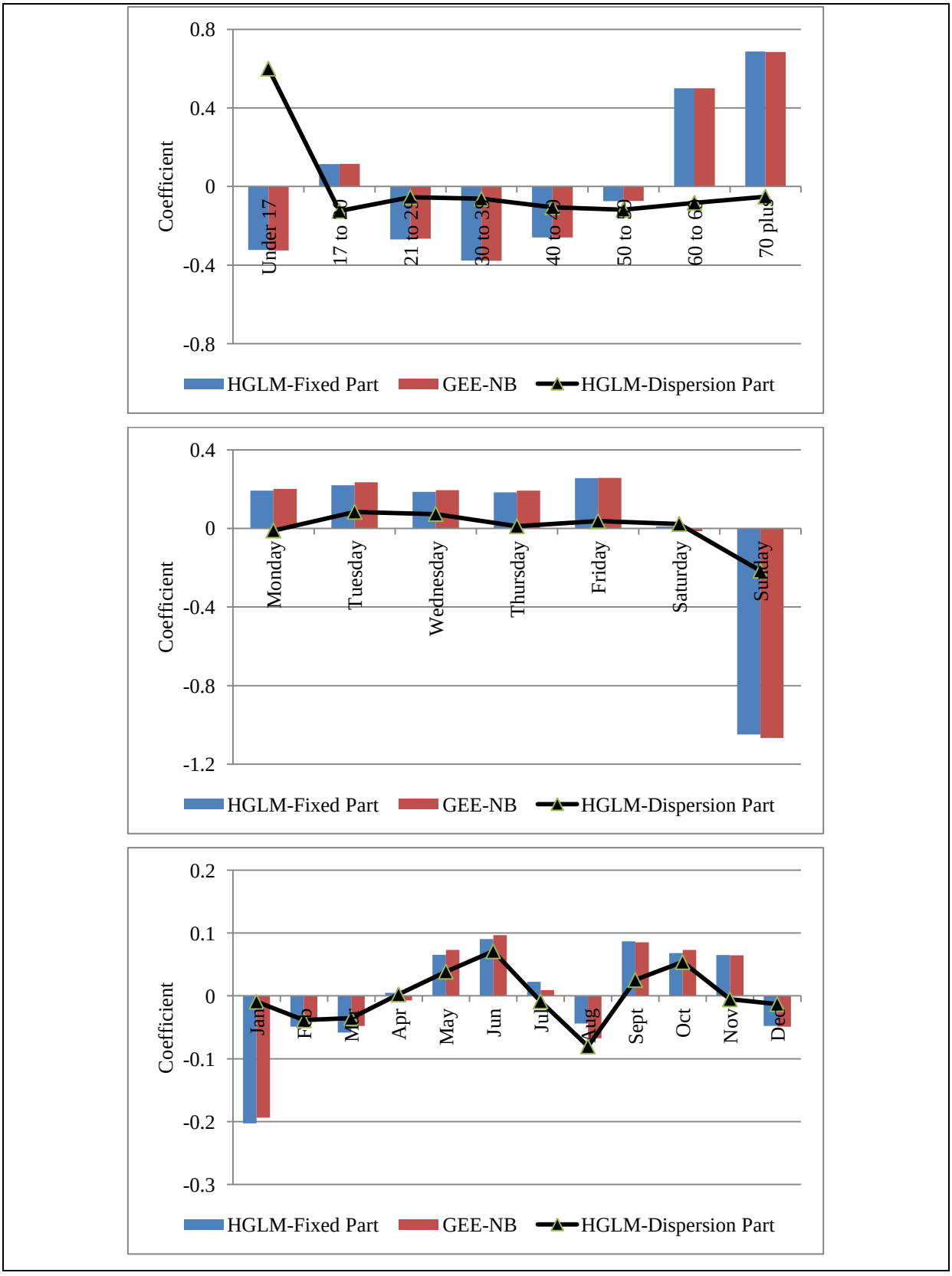


Appendix Table A5.12: Comparison of the coefficients and t values of the some variables by HGLM and GEE-AR1 (Bus: Dataset 9)

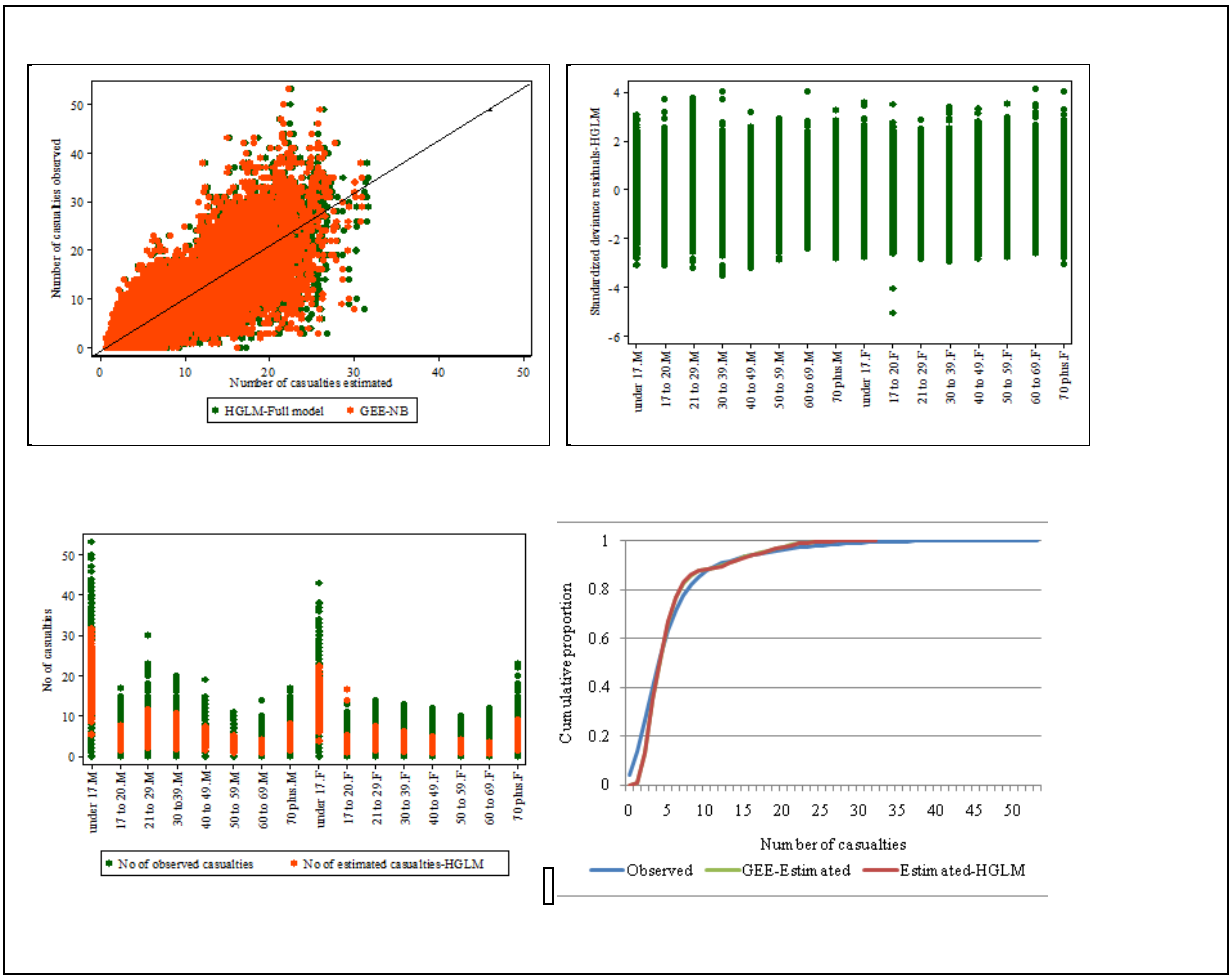
Variable (Bus)	**Coefficient HGLM**	***t* value HGLM**	**Coefficient GEE**	***t* value GEE**
Under 17	-0.323	-12.62	-0.325	-15.56
17-20	0.115	4.00	0.116	4.14
30-39	-0.377	-16.86	-0.377	-16.49
40-49	-0.259	-12.01	-0.259	-11.44
50-59	-0.074	-3.60	-0.074	-3.35
60-69	0.500	26.81	0.500	24.36
70 plus	0.687	44.94	0.685	37.50
Gender	-0.490	-36.52	-0.491	-36.58
Under 17.Male	0.208	5.52	0.208	6.72
17-20. Male	-0.199	-4.24	-0.208	-4.51
30-39. Male	0.417	13.02	0.421	12.64
40-49. Male	0.226	7.01	0.229	6.73
50-59. Male	-0.026	-0.79	-0.026	-0.74
60-69. Male	-0.483	-14.15	-0.482	-13.64
70 plus. Male	-0.331	-11.55	-0.325	-10.48
Monday	0.192	13.26	0.201	13.22
Tuesday	0.219	15.16	0.235	15.98
Wednesday	0.187	12.77	0.194	13.06
Thursday	0.184	12.91	0.192	12.90
Saturday	0.010	0.68	-0.014	-0.91
Sunday	-1.048	-47.75	-1.066	-47.41
January	-0.203	-5.48	-0.194	-8.72
February	-0.049	-1.35	-0.036	-1.65
March	-0.058	-1.62	-0.048	-2.28
April	0.005	0.13	-0.007	-0.35
May	0.065	1.82	0.073	3.57
June	0.090	2.52	0.096	4.72
July	0.022	0.63	0.009	0.45
August	-0.044	-1.23	-0.067	-3.17
September	0.087	2.43	0.085	4.14
October	0.068	1.90	0.073	3.60
December	-0.048	-1.31	-0.049	-2.24
Time	0.000	-6.91	0.000	-11.93
Holidays	-0.384	-12.87	-0.390	-13.73
New_Year	-0.446	-3.50	-0.448	-3.97
Christmas	-1.167	-6.25	-1.222	-6.36
Constant	-16.51	-73.37	-16.58	-74.91

Italics indicate the non-significant t values at the 5 percent level

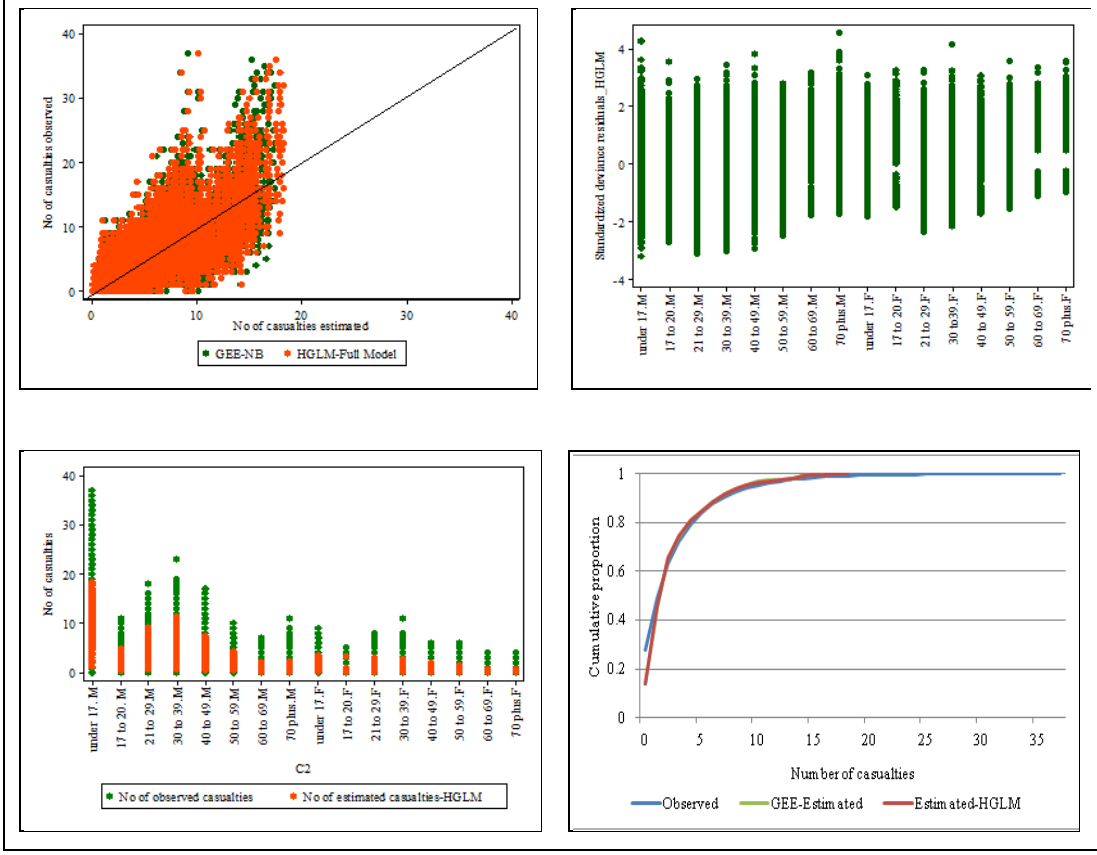
Appendix Figure A5.8: Comparison of coefficients by HGLM and GEE-AR1
(Bus: Dataset 9)



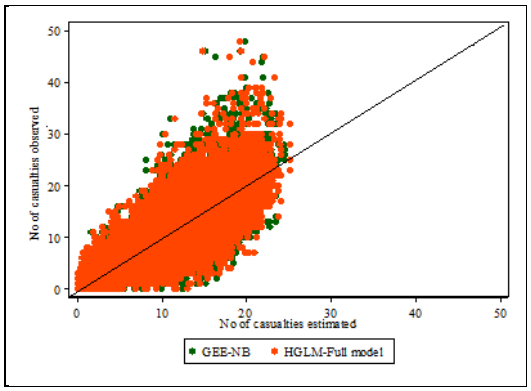
Appendix Figure A5.9: Comparison of casualties observed and estimated by HGLM and GEE-AR1, standardised deviance residuals, and cumulative proportion graphs (Walk)

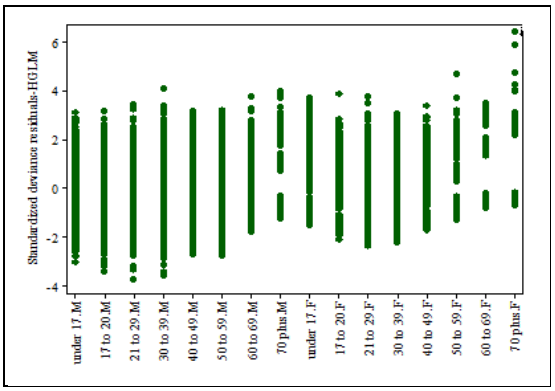


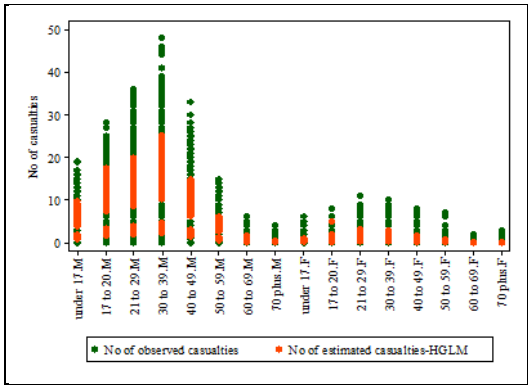
Appendix Figure A5.10: Comparison of casualties observed and estimated by HGLM and GEE-AR1, standardized deviance residuals and cumulative proportion graphs (Bicycle)

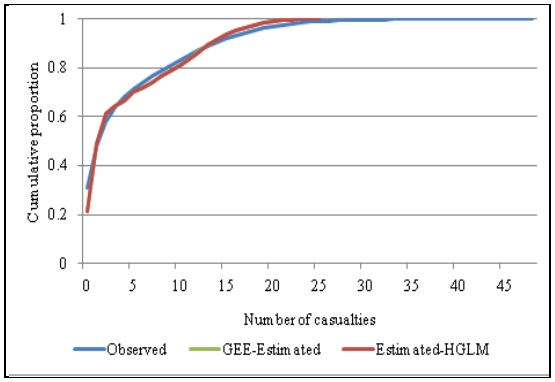


Appendix Figure A5.11: Comparison of casualties observed and estimated by HGLM and GEE-AR1, standardised deviance residuals and cumulative proportion graphs (Motorcycle)

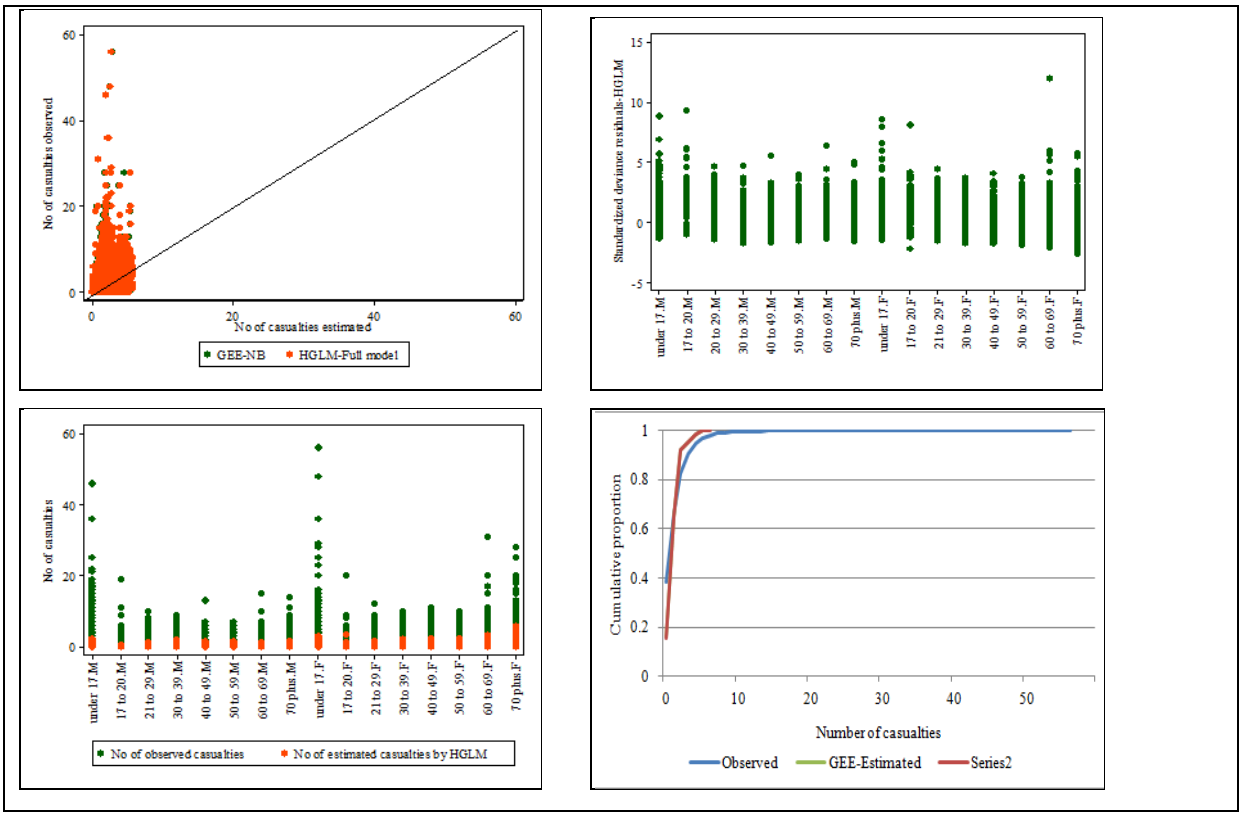




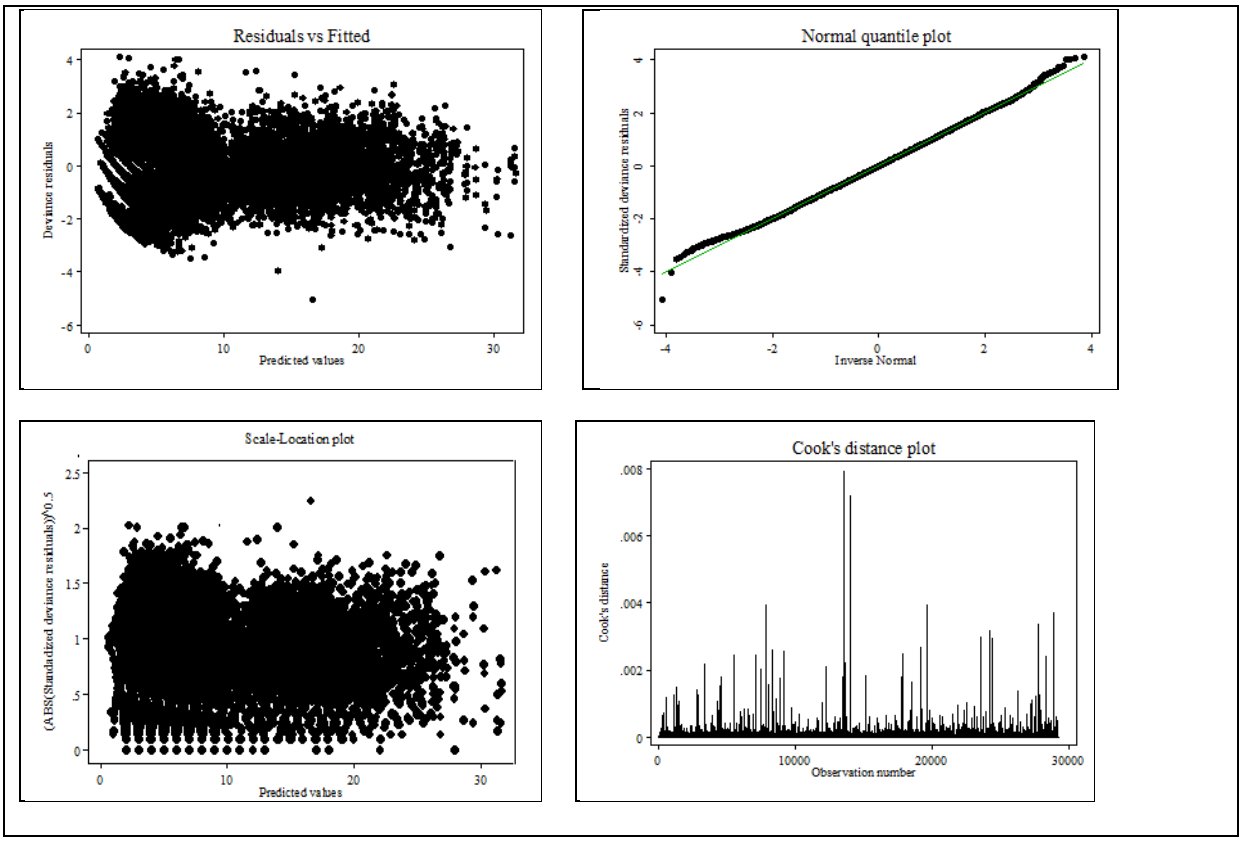




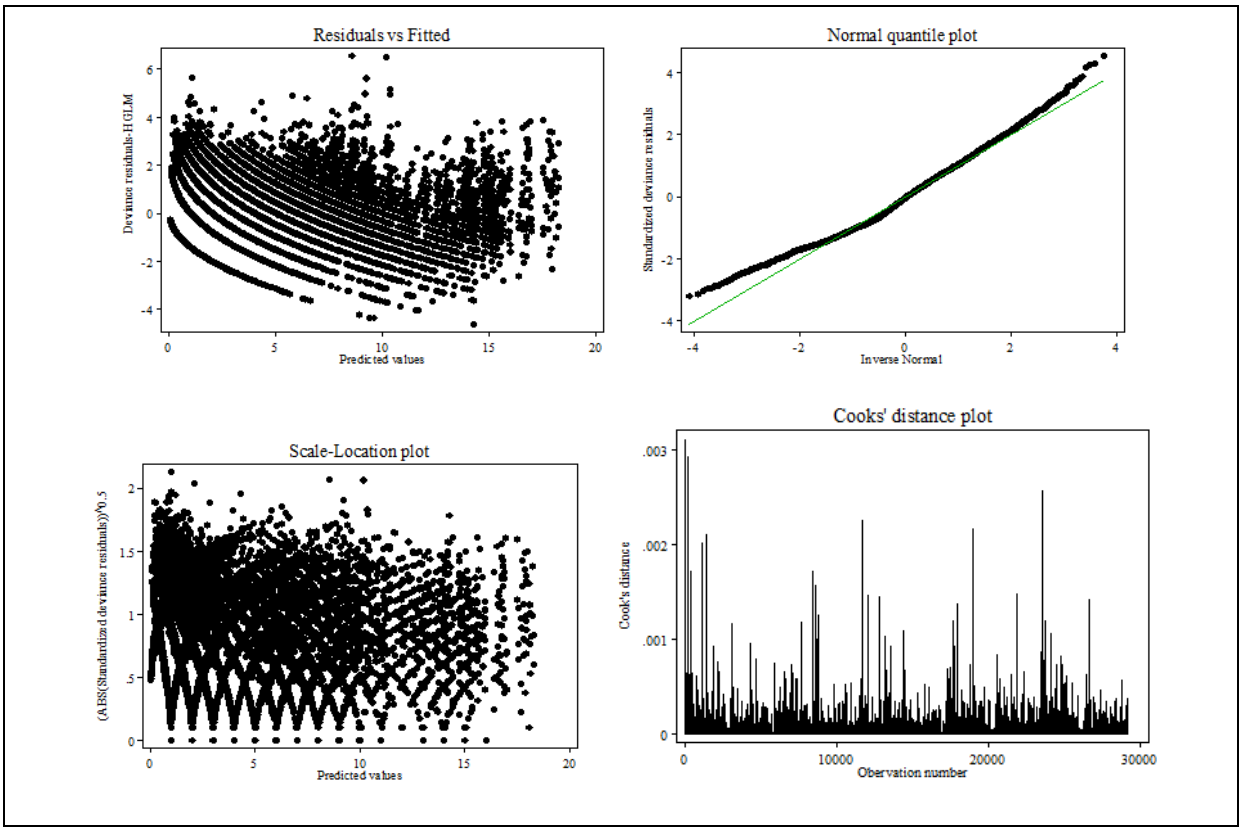
Appendix Figure A5.12: Comparison of casualties observed and estimated by HGLM and GEE-AR1, standardised deviance residuals and cumulative proportion graphs (Bus)



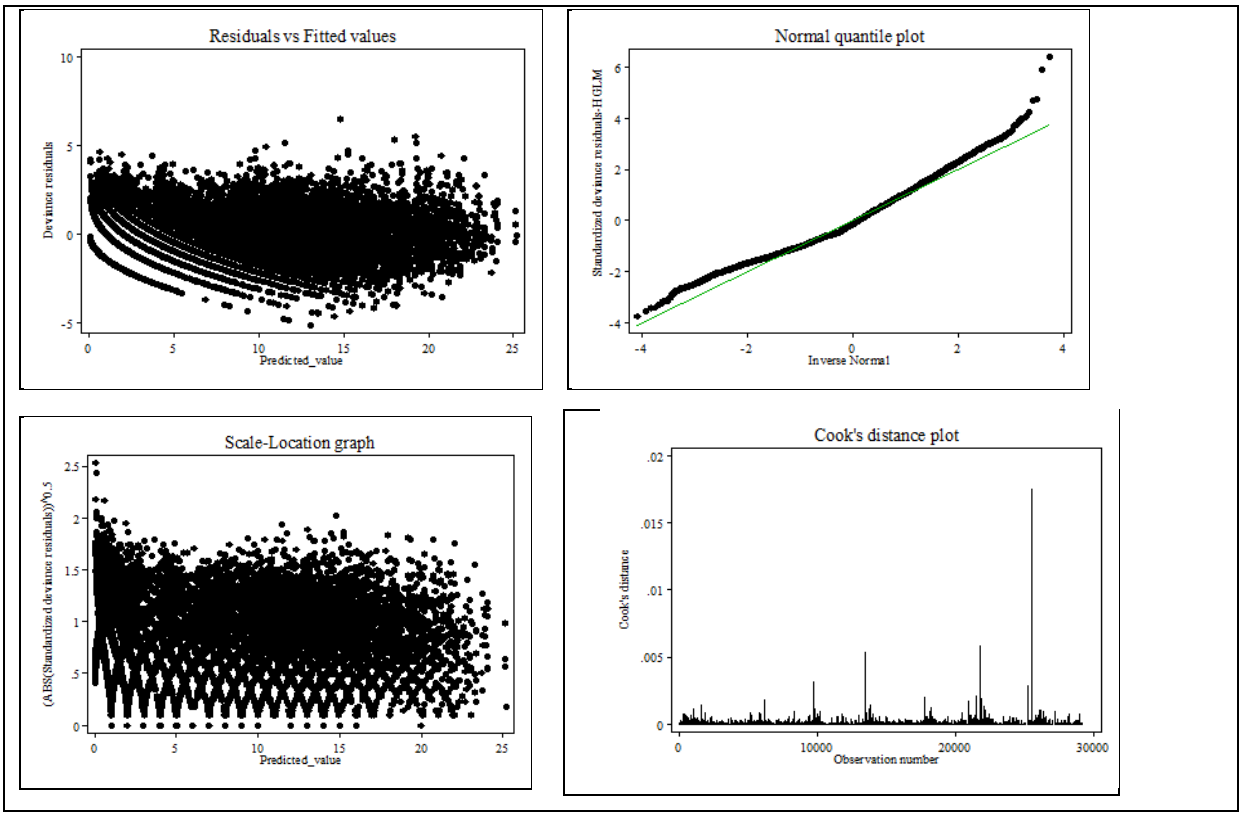
Appendix Figure A5.13: Diagnostic plot for Full model-HGLM-Walk



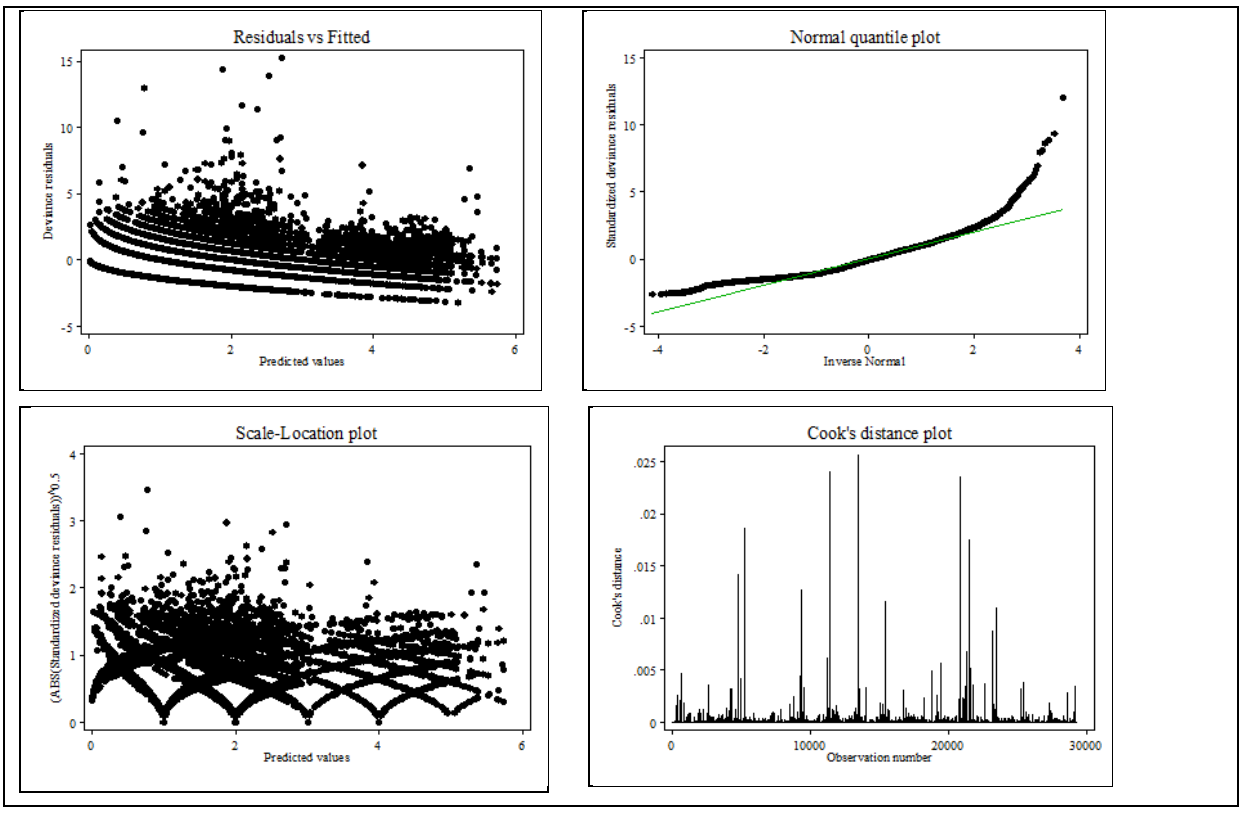
Appendix Figure A5.14: Diagnostic plot for Full model-HGLM-Bicycle



Appendix Figure A5.15: Diagnostic plot for Full model-HGLM-Motorcycle



Appendix Figure A5.16: Diagnostic plot for Full model-HGLM-Bus



Appendix Table A5.13: Results of Full model HGLM (Dataset 5: Car)

	Fixed part				Random Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Constant	-12.043	0.026	-464.97	Jan.2001	0.045	0.019	2.34
Under 17	-0.913	0.009	-100.18	Jan.2002	-0.028	0.019	-1.47
17-29	1.003	0.007	151.92	Jan.2003	-0.020	0.019	-1.04
30-39	0.420	0.005	79.70	Jan.2004	0.003	0.019	0.14
40-49	0.155	0.006	27.59	Jan.2005	-0.002	0.020	-0.11
50-59	-0.123	0.006	-20.81	Feb.2001	0.030	0.019	1.56
60-69	-0.467	0.007	-66.39	Feb.2002	0.032	0.019	1.73
70 plus	-0.834	0.007	-115.37	Feb.2003	0.013	0.019	0.70
Gender	0.023	0.002	13.11	Feb.2004	-0.070	0.019	-3.67
Under 17.Male	-0.280	0.013	-20.89	Feb.2005	-0.009	0.020	-0.44
17-20. Male	0.214	0.009	24.46	March.2001	-0.002	0.019	-0.11
30-39. Male	-0.016	0.007	-2.17	March.2002	0.007	0.019	0.37
40-49. Male	-0.075	0.008	-9.43	March.2003	-0.048	0.019	-2.55
50-59. Male	-0.131	0.009	-15.37	March.2004	0.041	0.019	2.17
60-69. Male	-0.013	0.010	-1.34	March.2005	0.001	0.020	0.02
70 plus. Male	0.240	0.010	23.22	April.2001	0.009	0.020	0.45
Monday	-0.038	0.004	-9.15	April.2002	-0.024	0.019	-1.26
Tuesday	-0.049	0.004	-12.50	April.2003	-0.011	0.019	-0.59
Wednesday	-0.025	0.004	-6.34	April.2004	0.027	0.019	1.45
Thursday	-0.006	0.004	-1.66	April.2005	-0.002	0.020	-0.09
Saturday	0.056	0.004	14.34	May.2001	-0.043	0.020	-2.21
Sunday	-0.084	0.004	-19.69	May.2002	0.003	0.019	0.18
January	0.006	0.016	0.40	May.2003	0.022	0.019	1.19
February	-0.018	0.016	-1.12	May.2004	-0.008	0.019	-0.42
March	-0.108	0.016	-6.87	May.2005	0.024	0.020	1.25
April	-0.079	0.016	-4.99	June.2001	-0.026	0.020	-1.33
May	-0.045	0.016	-2.84	June.2002	-0.022	0.019	-1.16
June	-0.047	0.016	-2.97	June.2003	0.007	0.019	0.38
July	-0.027	0.016	-1.70	June.2004	0.020	0.019	1.07
August	-0.031	0.016	-1.95	June.2005	0.020	0.020	1.01
September	-0.019	0.016	-1.21	July.2001	-0.020	0.020	-1.00
October	0.082	0.016	5.18	July.2002	0.028	0.019	1.50
December	0.143	0.016	9.02	July.2003	0.023	0.019	1.23
Time	0.000	0.000	-10.57	July.2004	-0.044	0.019	-2.32
Holidays	-0.057	0.006	-9.19	July.2005	0.011	0.020	0.56
New Year	-0.142	0.019	-7.61	Aug.2001	0.020	0.019	1.06
Christmas	-0.196	0.017	-11.85	Aug.2002	-0.028	0.019	-1.51
				Aug.2003	-0.050	0.019	-2.67
				Aug.2004	0.047	0.019	2.49
				Aug.2005	0.009	0.020	0.44

Continued-

	Random Part				Dispersion Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Sept.2001	0.013	0.019	0.68	Constant	0.802	0.008	96.71
Sept.2002	-0.053	0.019	-2.78	Under 17	0.428	0.022	19.51
Sept.2003	0.025	0.019	1.34	17 to 20	0.213	0.022	9.73
Sept.2004	0.021	0.019	1.09	30-39	0.190	0.022	8.67
Sept.2005	-0.008	0.020	-0.38	40-49	0.014	0.022	0.64
Oct.2001	-0.026	0.020	-1.30	50-59	-0.217	0.022	-9.88
Oct.2002	0.060	0.019	3.19	60-69	-0.455	0.022	-20.77
Oct.2003	-0.015	0.019	-0.82	70 plus	-0.382	0.022	-17.43
Oct.2004	0.015	0.019	0.78	Male	-0.015	0.008	-1.75
Oct.2005	-0.037	0.020	-1.84	Monday	0.018	0.020	0.91
Nov.2001	0.006	0.019	0.29	Tuesday	-0.057	0.020	-2.81
Nov.2002	0.042	0.018	2.31	Wednesday	-0.031	0.020	-1.54
Nov.2003	-0.007	0.018	-0.40	Thursday	-0.064	0.020	-3.14
Nov.2004	-0.057	0.019	-3.04	Saturday	0.014	0.020	0.69
Nov.2005	0.014	0.019	0.71	Sunday	0.089	0.020	4.37
Dec.2001	0.033	0.019	1.72	Jan	0.078	0.027	2.86
Dec.2002	0.003	0.019	0.13	Feb	-0.126	0.029	-4.44
Dec.2003	-0.043	0.019	-2.28	Mar	-0.091	0.027	-3.33
Dec.2004	0.011	0.019	0.56	Apr	-0.066	0.028	-2.38
Dec.2005	-0.005	0.020	-0.25	May	-0.041	0.027	-1.51
				Jun	0.057	0.028	2.07
				Jul	-0.015	0.027	-0.56
				Aug	-0.019	0.027	-0.69
				Sep	-0.022	0.028	-0.78
				Oct	0.131	0.027	4.82
				Dec	0.176	0.027	6.44

Appendix Table A5.14: Results of Full model HGLM (Dataset 6: Walk)

	Fixed part				Random Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Constant	-13.790	0.050	-274.47	Jan.2001	0.024	0.022	1.12
Under 17	0.670	0.010	65.35	Jan.2002	-0.046	0.022	-2.09
17-29	0.667	0.013	50.59	Jan.2003	0.002	0.022	0.09
30-39	-0.315	0.012	-26.55	Jan.2004	0.013	0.022	0.58
40-49	-0.445	0.013	-34.28	Jan.2005	0.006	0.022	0.26
50-59	-0.531	0.014	-38.50	Feb.2001	-0.001	0.022	-0.04
60-69	-0.401	0.015	-26.04	Feb.2002	0.027	0.021	1.29
70 plus	0.165	0.012	14.20	Feb.2003	-0.014	0.021	-0.64
Gender	0.176	0.003	53.97	Feb.2004	-0.002	0.022	-0.07
Under 17.Male	-0.052	0.014	-3.76	Feb.2005	-0.012	0.022	-0.54
17-20. Male	-0.005	0.018	-0.27	March.2001	-0.007	0.022	-0.34
30-39. Male	0.223	0.015	14.49	March.2002	0.023	0.021	1.10
40-49. Male	0.066	0.017	3.80	March.2003	-0.002	0.022	-0.10
50-59. Male	-0.078	0.019	-4.15	March.2004	0.002	0.022	0.10
60-69. Male	-0.169	0.022	-7.76	March.2005	-0.017	0.022	-0.74
70 plus. Male	-0.063	0.017	-3.67	April.2001	-0.009	0.022	-0.42
Monday	0.000	0.008	-0.02	April.2002	-0.016	0.022	-0.72
Tuesday	0.014	0.008	1.91	April.2003	-0.027	0.022	-1.20
Wednesday	0.019	0.007	2.50	April.2004	0.009	0.022	0.40
Thursday	0.058	0.007	7.88	April.2005	0.042	0.022	1.88
Saturday	0.049	0.008	6.43	May.2001	0.023	0.022	1.04
Sunday	-0.353	0.010	-36.62	May.2002	-0.015	0.022	-0.67
January	0.039	0.017	2.32	May.2003	-0.055	0.022	-2.48
February	0.018	0.016	1.08	May.2004	0.020	0.022	0.93
March	-0.032	0.017	-1.95	May.2005	0.025	0.022	1.10
April	-0.061	0.017	-3.62	June.2001	-0.004	0.022	-0.19
May	-0.028	0.017	-1.66	June.2002	-0.014	0.022	-0.63
June	-0.064	0.017	-3.81	June.2003	0.000	0.022	-0.02
July	-0.131	0.017	-7.73	June.2004	-0.007	0.022	-0.34
August	-0.167	0.017	-9.89	June.2005	0.025	0.022	1.14
September	0.002	0.017	0.12	July.2001	0.022	0.022	1.01
October	0.064	0.016	3.89	July.2002	-0.023	0.022	-1.03
December	0.173	0.017	10.21	July.2003	0.008	0.022	0.37
Time	0.000	0.000	-12.79	July.2004	-0.025	0.023	-1.11
Holidays	-0.135	0.013	-10.55	July.2005	0.016	0.023	0.72
New Year	0.057	0.032	1.79	Aug.2001	0.002	0.022	0.08
Christmas	-0.356	0.040	-8.96	Aug.2002	-0.007	0.022	-0.30
				Aug.2003	-0.003	0.022	-0.12
				Aug.2004	0.002	0.022	0.08
				Aug.2005	0.006	0.023	0.25

Continued-

	Random Part				Dispersion Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Sept.2001	0.004	0.022	0.19	Constant	0.361	0.008	43.61
Sept.2002	0.011	0.021	0.53	Under 17	0.724	0.022	33.02
Sept.2003	-0.007	0.022	-0.32	17 to 20	-0.079	0.022	-3.61
Sept.2004	0.017	0.022	0.77	30-39	-0.193	0.022	-8.81
Sept.2005	-0.026	0.022	-1.15	40-49	-0.190	0.022	-8.67
Oct.2001	0.008	0.022	0.39	50-59	-0.218	0.022	-9.94
Oct.2002	-0.006	0.021	-0.30	60-69	-0.143	0.022	-6.53
Oct.2003	-0.036	0.022	-1.69	70 plus	0.152	0.022	6.91
Oct.2004	0.048	0.021	2.23	Male	0.060	0.008	7.22
Oct.2005	-0.015	0.022	-0.67	Monday	-0.016	0.020	-0.77
Nov.2001	0.009	0.021	0.45	Tuesday	-0.027	0.020	-1.34
Nov.2002	0.016	0.021	0.78	Wednesday	-0.051	0.020	-2.50
Nov.2003	-0.017	0.021	-0.82	Thursday	-0.052	0.020	-2.58
Nov.2004	-0.029	0.021	-1.34	Saturday	0.059	0.020	2.91
Nov.2005	0.019	0.022	0.89	Sunday	0.173	0.020	8.51
Dec.2001	0.019	0.022	0.85	Jan	0.072	0.027	2.65
Dec.2002	0.005	0.022	0.22	Feb	-0.122	0.028	-4.30
Dec.2003	0.006	0.022	0.25	Mar	-0.055	0.027	-2.01
Dec.2004	-0.016	0.022	-0.71	Apr	0.011	0.028	0.38
Dec.2005	-0.014	0.023	-0.61	May	0.049	0.027	1.80
				Jun	-0.033	0.028	-1.20
				Jul	0.022	0.027	0.80
				Aug	-0.050	0.027	-1.84
				Sep	-0.069	0.028	-2.48
				Oct	0.000	0.027	-0.01
				Dec	0.212	0.027	7.77

Appendix Table A5.15: Results of Full model HGLM (Dataset 7: Cycle)

	Fixed part				Random Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Constant	-15.734	0.120	-131.15	Jan.2001	0.136	0.044	3.10
Under 17	0.286	0.022	13.28	Jan.2002	0.000	0.044	-0.01
17-29	0.512	0.027	18.89	Jan.2003	-0.099	0.045	-2.19
30-39	0.359	0.020	18.39	Jan.2004	-0.031	0.045	-0.69
40-49	0.044	0.023	1.93	Jan.2005	-0.021	0.047	-0.44
50-59	-0.168	0.025	-6.67	Feb.2001	0.083	0.045	1.84
60-69	-0.670	0.034	-19.89	Feb.2002	0.041	0.045	0.93
70 plus	-1.117	0.039	-29.03	Feb.2003	-0.025	0.045	-0.55
Gender	0.709	0.006	122.02	Feb.2004	-0.034	0.046	-0.74
Under 17.Male	0.213	0.024	8.85	Feb.2005	-0.073	0.048	-1.52
17-20. Male	0.109	0.031	3.58	March.2001	-0.047	0.045	-1.04
30-39. Male	0.025	0.022	1.12	March.2002	-0.005	0.044	-0.11
40-49. Male	0.013	0.026	0.48	March.2003	0.088	0.043	2.06
50-59. Male	-0.223	0.030	-7.55	March.2004	-0.002	0.044	-0.04
60-69. Male	-0.192	0.040	-4.83	March.2005	-0.040	0.046	-0.87
70 plus. Male	0.358	0.044	8.14	April.2001	-0.020	0.045	-0.43
Monday	0.131	0.010	12.87	April.2002	0.002	0.044	0.05
Tuesday	0.197	0.010	20.50	April.2003	0.009	0.043	0.22
Wednesday	0.204	0.010	21.16	April.2004	-0.026	0.044	-0.58
Thursday	0.183	0.010	18.73	April.2005	0.033	0.045	0.73
Saturday	-0.362	0.012	-29.50	May.2001	0.147	0.043	3.44
Sunday	-0.476	0.013	-36.68	May.2002	-0.091	0.044	-2.09
January	-0.162	0.036	-4.52	May.2003	-0.162	0.044	-3.67
February	-0.227	0.036	-6.31	May.2004	0.091	0.042	2.14
March	-0.228	0.035	-6.43	May.2005	-0.017	0.045	-0.38
April	-0.072	0.035	-2.05	June.2001	0.055	0.043	1.26
May	0.091	0.035	2.61	June.2002	-0.141	0.044	-3.19
June	0.206	0.035	5.92	June.2003	0.001	0.042	0.02
July	0.176	0.035	5.06	June.2004	0.038	0.042	0.90
August	0.157	0.035	4.51	June.2005	0.035	0.044	0.79
September	0.195	0.035	5.63	July.2001	0.077	0.043	1.79
October	0.078	0.035	2.23	July.2002	-0.086	0.043	-2.00
December	-0.234	0.037	-6.40	July.2003	0.000	0.042	0.00
Time	0.000	0.000	-4.89	July.2004	-0.056	0.043	-1.30
Holidays	-0.235	0.018	-12.86	July.2005	0.055	0.044	1.27
New Year	-0.582	0.089	-6.55	Aug.2001	0.032	0.043	0.74
Christmas	-0.523	0.080	-6.53	Aug.2002	-0.096	0.043	-2.22
				Aug.2003	0.003	0.042	0.07
				Aug.2004	-0.027	0.043	-0.62
				Aug.2005	0.079	0.043	1.83

Continued-

	Random Part				Dispersion Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Sept.2001	-0.103	0.045	-2.32	Constant	0.238	0.008	28.7
Sept.2002	0.027	0.042	0.65	Under 17	0.457	0.022	20.83
Sept.2003	-0.012	0.042	-0.28	17 to 20	-0.159	0.022	-7.27
Sept.2004	0.035	0.042	0.84	30-39	-0.015	0.022	-0.67
Sept.2005	0.045	0.043	1.05	40-49	0.011	0.022	0.49
Oct.2001	-0.014	0.044	-0.31	50-59	-0.055	0.022	-2.5
Oct.2002	-0.010	0.042	-0.24	60-69	-0.194	0.022	-8.84
Oct.2003	-0.004	0.042	-0.09	70 plus	0.030	0.022	1.39
Oct.2004	-0.024	0.043	-0.55	Male	0.108	0.008	13.02
Oct.2005	0.050	0.044	1.14	Monday	-0.006	0.020	-0.27
Nov.2001	0.010	0.044	0.23	Tuesday	-0.020	0.020	-0.96
Nov.2002	-0.009	0.043	-0.21	Wednesday	0.001	0.020	0.04
Nov.2003	-0.036	0.043	-0.82	Thursday	0.008	0.020	0.39
Nov.2004	-0.040	0.044	-0.91	Saturday	0.013	0.020	0.64
Nov.2005	0.071	0.044	1.60	Sunday	0.019	0.020	0.93
Dec.2001	0.017	0.047	0.36	Jan	0.015	0.027	0.54
Dec.2002	-0.055	0.047	-1.17	Feb	-0.015	0.028	-0.54
Dec.2003	-0.049	0.047	-1.06	Mar	-0.126	0.027	-4.63
Dec.2004	0.025	0.046	0.54	Apr	-0.041	0.028	-1.47
Dec.2005	0.058	0.047	1.23	May	0.030	0.027	1.11
				Jun	0.057	0.028	2.04
				Jul	0.043	0.027	1.57
				Aug	0.005	0.027	0.2
				Sep	-0.014	0.028	-0.49
				Oct	-0.064	0.027	-2.35
				Dec	0.126	0.027	4.61

Appendix Table A5.16: Results of Full model HGLM (Dataset 8: Motorcycle)

	Fixed part				Random Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Constant	-15.685	0.088	-177.42	Jan.2001	0.042	0.039	1.07
Under 17	-0.328	0.027	-12.26	Jan.2002	-0.070	0.039	-1.80
17-29	1.655	0.022	74.96	Jan.2003	-0.005	0.038	-0.12
30-39	0.903	0.021	44.06	Jan.2004	-0.046	0.039	-1.18
40-49	0.389	0.025	15.77	Jan.2005	0.072	0.039	1.84
50-59	-0.294	0.031	-9.58	Feb.2001	0.035	0.039	0.91
60-69	-1.218	0.046	-26.76	Feb.2002	0.014	0.038	0.38
70 plus	-2.338	0.054	-43.51	Feb.2003	0.037	0.038	0.97
Gender	1.036	0.007	150.91	Feb.2004	-0.027	0.039	-0.68
Under 17.Male	0.000	0.029	-0.01	Feb.2005	-0.064	0.041	-1.56
17-20. Male	0.042	0.024	1.75	March.2001	-0.116	0.040	-2.93
30-39. Male	0.085	0.023	3.78	March.2002	0.056	0.037	1.51
40-49. Male	0.151	0.027	5.61	March.2003	0.157	0.036	4.36
50-59. Male	0.047	0.034	1.41	March.2004	-0.058	0.039	-1.51
60-69. Male	-0.016	0.050	-0.31	March.2005	-0.064	0.040	-1.61
70 plus. Male	-0.127	0.062	-2.06	April.2001	-0.117	0.039	-3.00
Monday	-0.014	0.009	-1.60	April.2002	0.092	0.036	2.53
Tuesday	-0.002	0.009	-0.22	April.2003	0.028	0.037	0.76
Wednesday	0.051	0.008	6.15	April.2004	-0.007	0.037	-0.19
Thursday	0.021	0.009	2.47	April.2005	-0.007	0.039	-0.17
Saturday	-0.072	0.009	-8.07	May.2001	0.042	0.037	1.14
Sunday	-0.081	0.010	-8.18	May.2002	-0.058	0.037	-1.57
January	-0.311	0.031	-10.10	May.2003	-0.067	0.037	-1.83
February	-0.268	0.031	-8.71	May.2004	0.047	0.036	1.28
March	-0.139	0.030	-4.57	May.2005	0.029	0.038	0.78
April	-0.002	0.030	-0.07	June.2001	-0.037	0.038	-0.96
May	0.121	0.030	4.05	June.2002	-0.050	0.037	-1.34
June	0.165	0.030	5.51	June.2003	0.061	0.036	1.69
July	0.141	0.030	4.71	June.2004	-0.002	0.037	-0.05
August	0.149	0.030	5.01	June.2005	0.023	0.038	0.61
September	0.218	0.030	7.34	July.2001	-0.005	0.037	-0.13
October	0.113	0.030	3.79	July.2002	-0.017	0.036	-0.48
December	-0.232	0.031	-7.49	July.2003	0.040	0.036	1.13
Time	0.000	0.000	-7.08	July.2004	-0.036	0.037	-0.97
Holidays	0.030	0.012	2.46	July.2005	0.016	0.038	0.42
New Year	-0.566	0.060	-9.47	Aug.2001	-0.049	0.037	-1.30
Christmas	-0.779	0.064	-12.25	Aug.2002	-0.028	0.036	-0.77
				Aug.2003	0.102	0.035	2.94
				Aug.2004	-0.028	0.036	-0.78
				Aug.2005	-0.005	0.038	-0.13

Continued-

	Random Part				Dispersion Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Sept.2001	-0.085	0.038	-2.25	Constant	0.179	0.008	21.59
Sept.2002	0.052	0.035	1.46	Under 17	0.106	0.022	4.83
Sept.2003	0.050	0.035	1.41	17 to 20	0.165	0.022	7.53
Sept.2004	0.003	0.036	0.08	30-39	0.304	0.022	13.84
Sept.2005	-0.026	0.038	-0.69	40-49	0.244	0.022	11.12
Oct.2001	-0.024	0.037	-0.63	50-59	0.027	0.022	1.24
Oct.2002	0.009	0.036	0.26	60-69	-0.310	0.022	-14.15
Oct.2003	0.071	0.035	2.00	70 plus	-0.688	0.022	-31.41
Oct.2004	-0.062	0.037	-1.68	Male	0.209	0.008	25.18
Oct.2005	0.001	0.038	0.02	Monday	-0.049	0.020	-2.42
Nov.2001	0.033	0.037	0.88	Tuesday	-0.046	0.020	-2.26
Nov.2002	0.000	0.037	0.01	Wednesday	-0.051	0.020	-2.51
Nov.2003	-0.033	0.037	-0.90	Thursday	-0.026	0.020	-1.30
Nov.2004	0.004	0.037	0.11	Saturday	-0.006	0.020	-0.28
Nov.2005	-0.005	0.038	-0.14	Sunday	0.225	0.020	11.07
Dec.2001	0.027	0.040	0.68	Jan	-0.060	0.027	-2.18
Dec.2002	0.003	0.039	0.07	Feb	-0.079	0.028	-2.78
Dec.2003	0.004	0.039	0.10	Mar	0.010	0.027	0.37
Dec.2004	0.026	0.039	0.67	Apr	0.017	0.028	0.60
Dec.2005	-0.062	0.041	-1.51	May	0.026	0.027	0.96
				Jun	0.123	0.028	4.46
				Jul	0.030	0.027	1.08
				Aug	-0.032	0.027	-1.17
				Sep	-0.018	0.028	-0.64
				Oct	-0.038	0.027	-1.38
				Dec	0.055	0.027	2.03

Appendix Table A5.17: Results of Full model HGLM (Dataset 9: Bus)

	Fixed part				Random Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Constant	-16.515	0.225	-73.37	Jan.2001	0.007	0.048	0.15
Under 17	-0.323	0.026	-12.62	Jan.2002	-0.098	0.050	-1.96
17-29	0.115	0.029	4.00	Jan.2003	-0.044	0.049	-0.89
30-39	-0.377	0.022	-16.86	Jan.2004	0.099	0.047	2.10
40-49	-0.259	0.022	-12.01	Jan.2005	0.025	0.050	0.50
50-59	-0.074	0.021	-3.60	Feb.2001	0.004	0.048	0.09
60-69	0.500	0.019	26.81	Feb.2002	0.033	0.047	0.71
70 plus	0.687	0.015	44.94	Feb.2003	-0.072	0.048	-1.48
Gender	-0.490	0.013	-36.52	Feb.2004	0.100	0.046	2.17
Under 17.Male	0.208	0.038	5.52	Feb.2005	-0.077	0.050	-1.53
17-20. Male	-0.199	0.047	-4.24	March.2001	-0.028	0.048	-0.58
30-39. Male	0.417	0.032	13.02	March.2002	0.041	0.046	0.89
40-49. Male	0.226	0.032	7.01	March.2003	-0.056	0.048	-1.17
50-59. Male	-0.026	0.033	-0.79	March.2004	0.065	0.046	1.41
60-69. Male	-0.483	0.034	-14.15	March.2005	-0.028	0.049	-0.56
70 plus. Male	-0.331	0.029	-11.55	April.2001	0.035	0.047	0.75
Monday	0.192	0.015	13.26	April.2002	-0.067	0.048	-1.40
Tuesday	0.219	0.015	15.16	April.2003	0.008	0.047	0.17
Wednesday	0.187	0.015	12.77	April.2004	0.030	0.047	0.64
Thursday	0.184	0.014	12.91	April.2005	-0.010	0.049	-0.20
Saturday	0.010	0.015	0.68	May.2001	0.001	0.047	0.02
Sunday	-1.048	0.022	-47.75	May.2002	-0.100	0.048	-2.09
January	-0.203	0.037	-5.48	May.2003	-0.002	0.046	-0.03
February	-0.049	0.036	-1.35	May.2004	0.057	0.046	1.23
March	-0.058	0.036	-1.62	May.2005	0.037	0.048	0.77
April	0.005	0.036	0.13	June.2001	0.033	0.047	0.70
May	0.065	0.036	1.82	June.2002	-0.056	0.048	-1.17
June	0.090	0.036	2.52	June.2003	-0.007	0.047	-0.16
July	0.022	0.036	0.63	June.2004	0.017	0.047	0.35
August	-0.044	0.036	-1.23	June.2005	0.012	0.048	0.25
September	0.087	0.036	2.43	July.2001	0.020	0.047	0.43
October	0.068	0.036	1.90	July.2002	-0.011	0.046	-0.23
December	-0.048	0.037	-1.31	July.2003	0.005	0.046	0.10
Time	0.000	0.000	-6.91	July.2004	-0.027	0.047	-0.57
Holidays	-0.384	0.030	-12.87	July.2005	0.012	0.048	0.26
New Year	-0.446	0.128	-3.50	Aug.2001	-0.005	0.047	-0.10
Christmas	-1.167	0.187	-6.25	Aug.2002	-0.004	0.046	-0.09
				Aug.2003	0.010	0.046	0.21
				Aug.2004	0.020	0.047	0.43
				Aug.2005	-0.021	0.048	-0.43

Continued-

	Random Part				Dispersion Part		
Variable	Coefficient	S.E	*t* value	Coefficient	Coefficient	S.E	*t* value
Sept.2001	-0.020	0.047	-0.43	Constant	0.405	0.012	34.53
Sept.2002	-0.003	0.046	-0.05	Under 17	0.600	0.022	27.37
Sept.2003	0.164	0.044	3.75	17 to 20	-0.125	0.022	-5.69
Sept.2004	-0.070	0.048	-1.46	30-39	-0.062	0.022	-2.84
Sept.2005	-0.092	0.050	-1.86	40-49	-0.106	0.022	-4.81
Oct.2001	-0.002	0.047	-0.04	50-59	-0.119	0.022	-5.41
Oct.2002	0.017	0.046	0.37	60-69	-0.083	0.022	-3.80
Oct.2003	0.050	0.046	1.10	70 plus	-0.051	0.022	-2.34
Oct.2004	-0.077	0.048	-1.60	Male	-0.109	0.017	-6.55
Oct.2005	0.007	0.048	0.15	Monday	-0.011	0.020	-0.53
Nov.2001	-0.034	0.048	-0.71	Tuesday	0.084	0.020	4.13
Nov.2002	-0.032	0.047	-0.69	Wednesday	0.073	0.020	3.59
Nov.2003	-0.016	0.047	-0.34	Thursday	0.011	0.020	0.53
Nov.2004	0.053	0.046	1.16	Saturday	0.022	0.020	1.10
Nov.2005	0.025	0.048	0.53	Sunday	-0.216	0.020	-10.64
Dec.2001	0.065	0.047	1.39	Jan	-0.009	0.027	-0.34
Dec.2002	0.042	0.047	0.90	Feb	-0.039	0.028	-1.36
Dec.2003	-0.052	0.048	-1.07	Mar	-0.035	0.027	-1.30
Dec.2004	0.006	0.048	0.13	Apr	0.002	0.028	0.08
Dec.2005	-0.069	0.050	-1.36	May	0.039	0.027	1.41
				Jun	0.071	0.028	2.57
				Jul	-0.009	0.027	-0.32
				Aug	-0.080	0.027	-2.95
				Sep	0.025	0.028	0.91
				Oct	0.054	0.027	1.97
				Dec	-0.013	0.027	-0.48