2010
2011

BEDRIJFSECONOMISCHE WETENSCHAPPEN

master in de verkeerskunde: verkeersveiligheid

Masterproef

Cluster analysis of crashes on intersection types in Belgium

Promotor :
Prof. dr. Tom BRIJS
dr. Benoit DEPAIRE

John Ediebah

Masterproef voorgedragen tot het bekomen van de graad van master in de verkeerskunde , afstudeerrichting verkeersveiligheid

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

BEDRIJFSECONOMISCHE WETENSCHAPPEN

master in de verkeerskunde: verkeersveiligheid

Masterproef

Cluster analysis of crashes on intersection types in Belgium

Promotor :
Prof. dr. Tom BRIJS
dr. Benoit DEPAIRE

John Ediebah

Masterproef voorgedragen tot het bekomen van de graad van master in de verkeerskunde , afstudeerrichting verkeersveiligheid

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Preface

Research studies focusing on factors leading to road crashes necessitate the collection of relevant data regarding several aspects including the road user, vehicle, road network and the environment among others. Due to the complexity of road crashes, the data collected is usually very large, multidimensional and highly heterogeneous. Therefore, it becomes primordial to reduce this heterogeneity in order to retrieve hidden information. Moreover, road intersections play a critical role in the safety and mobility performance of a road network. This report which culminates with the completion of my master studies at the University of Hasselt seeks to identify dominant crash or accident types at intersections in Belgium in order to promote road safety.

I hereby extend my gratitude to those people who contributed to the realization of this project in particular and to the successful completion of my studies in general. Sincere thanks to my promoter, Prof. dr. Tom Brijs and co-promoter, Dr. Benoit Depaire for their constant guidance and constructive criticisms. My profound gratitude to Mevrouw Nadine Smeyers, Mevrouw Isabel Thys and the entire staff at the student secretariat for their administrative assistance.

My studies would have been impossible without the invaluable support of my brother, Ediebah Divine Ewane, my beloved mother, Ediebah Roseline Mboulle, my sister Ediebah Ebude Vivian and my other brother and sisters for their words of encouragement and patience.

Finally, kind regards to all my classmates and friends especially Goele Lipkens who contributed (or are contributing) to make my stay in Belgium a unique experience.

EDIEBAH John Edie

Hasselt, Belgium

July 20th, 2011.

Summary

A great number of road traffic crashes or accidents occur at intersections worldwide. In the year 2005, about 34% of road crashes in Belgium occurred at intersections alone. Due to the complexity of road crashes, the crash data collected is usually very large and heterogeneous. Data mining techniques have become increasingly useful to analyse this type of data in order to reduce its heterogeneity and discover obscure patterns. The main objective of this study is to determine the dominant crash or accident types at various types of intersections in Belgium using cluster analysis. The purpose of cluster analysis is to group similar observations or objects; in this case crashes or accidents into homogenous groups or clusters from which meaningful insights can be obtained. Distance-based clustering techniques including K-means and hierarchical clustering and also fuzzy clustering are first reviewed and their strengths and weaknesses highlighted. These traditional distance-based clustering methods are usually ad hoc and not statistically based thereby producing less reliable results. The model-based clustering technique known as latent class cluster analysis (LCCA) or finite mixture model is the main analytical tool in this research. A pre-analysis using basic exploratory data analysis (EDA) is first conducted for all the crashes that occurred at intersections in the year 2005. The crash data is then segmented according to the different types of intersections. In all, six intersections are distinguished based on the type of traffic control. Three of the data sets representing about 98% of the overall crashes at intersections (intersection with right-of-way sign B1 or B5, intersection with right-of-way to traffic from right and intersection with functioning traffic lights) are then clustered individually using latent class cluster analysis with the aid of the Latent Gold software package. The three other data sets representing the other intersection types (intersection with a traffic policeman, intersection with a defective traffic light and right-of-way to traffic from the right and intersection with a defective traffic light and right-of-way sign B1 or B5) could not be effectively modeled due to an insufficient number of crashes.

From the pre-analysis, it is shown that about 52% of victims at intersections were involved in crashes at intersections with a right-of-way sign B1 (yield) or B5 (stop) sign. Surprisingly, the highest fatality rate was registered at intersections with a traffic policeman. However, only 0.45% of crashes occurred at this intersection type. Furthermore, the most common type of collision was side-collision between drivers

making up 57% and in terms of crash severity, collision with an obstacle (on and off the road) has the highest fatality rate while rear-end collisions has the least. Results from LCCA show that a significant number of crashes at the various intersections involved violation of right of way/priority or traffic light. The accident share increases as the intersection type moved from signal-controlled (Intersec_TL) to right-of-way signs (Intersec_RS) and finally to uncontrolled (Intersec_TR) intersection. Hence, signalised intersections are better in terms of road safety compared to the other two as they minimize the possibility of violation which is a key cause of accidents at intersections. In brief, Latent class cluster analysis is an effective data segmentation tool which assists in reducing the heterogeneity of crash data by identifying homogeneous groups which can facilitate the discovery of hidden patterns. Moreover, it facilitates the implementation of more in-depth analytical tools using predictive models such as injury risk analysis which focuses on the outcome of crashes.

Keywords: Intersections, road crashes, heterogeneous, homogeneous, cluster analysis, K-means, hierarchical clustering, cluster, segmentation, exploratory data analysis (EDA), latent class cluster analysis (LCCA), Finite mixture model.

Dedication

To my father, EDIEBAH Hilary Simon, of blessed memory.

Table of Contents

p.

Preface i

Summary ii

Dedication iV

Table of Contents V

List of Tables ix

List of Figures x

Chapter 1: INTRODUCTION - 1 -

1.1 Background - 1 -

1.2 General Aspects of Road Safety - 2 -

1.3 Intersections and Traffic Safety - 4 -

1.3.1 Road Safety at Intersections - 4 -

1.3.2 Types of Intersections - 6 -

1.4 Road Safety and Infrastructure in Belgium - 7 -

1.4.1 Overview of Road Safety - 7 -

1.4.2 The Road Network - 9 -

1.5 Scope of Study and Research Objectives - 13 -

1.6 Research Method - 15 -

1.7 Structure of the Report - 15 -

Chapter 2: STANDARD CLUSTER ANALYSIS - 17 -

2.1 Definition of Cluster Analysis - 19 -

2.2 Assessing Similarity and Dissimilarity in Cluster Analysis - 22 -

2.3 Selected Distance Functions between Patterns x and y - 22 -

2.3.1 Distance Metrics for Continuous Variables - 23 -

2.3.2 Distance Measures for Binary Variables - 28 -

2.4 Types of Clustering Algorithms - 30 -

2.4.1 Hierarchical Clustering - 30 -

2.4.1a Agglomerative or Bottom-up Technique - 31 -

2.4.1b Divisive or Top-down Technique - 31 -

2.5 Types of Linkage Functions - 32 -

2.6 Advantages and Disadvantages of Hierarchical Clustering - 37 -

2.6.1 Advantages - 37 -

2.6.2 Disadvantages - 37 -

2.7 Partitional of Non-hierarchical Clustering - 38 -

2.7.1 K-means Clustering - 38 -

2.7.1a Advantages - 39 -

2.7.1b Disadvantages - 39 -

2.7.2 Fuzzy C-means (FCM) Clustering - 40 -

Chapter 3: LATENT CLASS MODELS (LCM) AND LATENT CLASS CLUSTER ANALYSIS (LCCA) - 43 -

3.1 Latent Class Models (LCM) - 43 -

3.2 Latent Class Cluster Analysis (LCCA) - 45 -

3.2.1 LCCA for Continuous Variables - 46 -

3.2.2 LCCA for Mixed-Mode Indicators - 49 -

3.2.3 Inclusion of Covariates - 51 -

3.2.4 Model Estimation - 52 -

3.2.5 Model Selection - 53 -

Chapter 4: METHODOLOGY - 57 -

4.1 Data Description and Preparation - 57 -

4.1.1 Other Selected Variables - 62 -

4.1.2 Data Segmentation - 65 -

4.2 Research Method - 65 -

4.2.1 Latent Class Cluster Analysis (LCCA) - 66 -

4.2.2 Advantages of LCCA over Traditional Clustering Methods - 66 -

4.2.3 Implementation of LCCA in Latent Gold and Data Processing - 67 -

Chapter 5: ANALYSIS OF RESULTS AND DISCUSSION - 71 -

5.1 Results of Basic Exploratory Data Analysis (EDA) - 71 -

5.1.1 Crashes at Intersection Types - 71 -

5.1.2 Collision Types (*COL_T*) - 74 -

5.1.3 Crashes according to time of the day (*HOUR*) - 76 -

5.1.4 Crashes during the Week-end - 78 -

5.1.5 Seasonal distribution of crashes - 78 -

5.1.6 Crashes according to built-up area (*BUA*) - 79 -

5.2 Additional description of Intersec_TP and Intersec_DTLTR - 80 -

5.2.1 Intersec_TP - 80 -

5.2.2 Intersec_DTLTR - 81 -

5.3 Results of LCCA in Latent Gold - 81 -

5.3.1 Intersection with Functioning Three-colored Traffic Lights (Intersec_TL) - 82 -

5.3.2 Intersection with the Right-of-way sign B1 or B5 present (Intersec_RS) - 84 -

5.3.3 Intersection with Right of way to Traffic from the right (Intersec_TR) - 87 -

5.4 Discussion and Summary - 88 -

5.4.1 Interpretation of Results of Pre-analysis using Descriptive Statistics - 88 -

5.4.2 Summary of crash types - 91 -

5.4.3 Comparison of Crash types at various Intersections - 92 -

Chapter 6: CONCLUSION AND FURTHER RESEARCH - 95 -

6.1 Conclusion and Findings - 95 -

6.2 Limitations and Further Research - 97 -

Bibliography - 99 -

Appendices - 109 -

Appendix 1: SAS codes - 109 -

Appendix 2: Latent Gold Output - 110 -

Appendix 3: Additional descriptive statistics for Intersec_TP - 117 -

Appendix 4: Additional descriptive statistics for Intersec_DTLTR -120 -

List of Tables

Table 1-1: Traffic accidents and casualties in Belgium - 8 -

Table 1-2: The Road Network in Belgium (2005) - 10 -

Table 1-3: Motorization level in Belgium - 12 -

Table 4-1a: Traffic Control at Intersection (*TRAFFIC_C*) - 59 -

Table 4-1b: Adjusted Traffic control at Intersection - 60 -

Table 4-2: Time of the day (*HOUR*) - 60 -

Table 4-3: Aggregation of months into Seasons (*SEASON*) - 61 -

Table 4-4: Adjusted categories for road conditions (Road_C) - 61 -

Table 4-5: Adjusted categories for Weather conditions (*Weather*) - 62 -

Table 4-6: List of Selected Variables - 64 -

Table 4-7: Sample data sets based on Intersection type - 65 -

Table 5-1: Frequency distribution of crashes at various intersections - 71 -

Table 5-2: Severity of crashes based on Intersection type - 73 -

Table 5-3: Frequency distribution of collision types (COL_T) - 74 -

Table 5-4: Collision type and Severity of Crashes - 75 -

Table 5-5: Number of crashes according to time of the day (*HOUR*) - 77 -

Table 5-6: Number of crashes during the week-end - 78 -

Table 5-7: Seasonal distribution of crashes (*SEASON*) - 79 -

Table 5-8: Number of crashes according to built-up area (BUA) - 80 -

Table 5-9: Crash types at Intersec_TL - 84 -

Table 5-10: Crash types at Intersec_RS - 86 -

Table 5-11: Crash types at Intersec_TR - 88 -

Table 5-12: Summary of Crash types at various Intersections - 92 -

List of Figures

Figure 1-1: Evolution of road safety in Europe (1991 – 2009) - 3 -

Figure 1-2: Increasing car use in Europe compared to other modes - 4 -

Figure 1-3: A T-intersection showing the different conflict points - 6 -

Figure 1-4: Evolution of traffic fatalities in Belgium (1991-2006) - 9 -

Figure 1-5: A-road system with major R-roads in Belgium - 11 -

Figure 1-6: Passenger car population in Belgium (2009) - 13 -

Figure 2-1: Stages in clustering - 21 -

Figure 2-2: Distinction between Manhattan and Euclidean distance metrics - 25 -

Figure 2-3: Examples of distance functions - 28 -

Figure 2-4: A Dendogram - 31 -

Figure 2-5: Hierarchical clustering - 32 -

Figure 2-6: Two clusters and three main ways of computing the distance between them - 35 -

Figure 4-1: Road crashes at Intersections in Belgium - 58 -

Figure 4-2: Screenshot of the Variable tab in Latent Gold - 69 -

Figure 5-1: Proportion of crashes at various intersection types - 72 -

Figure 5-2: Severity of crashes based on intersection types - 73 -

Figure 5-3: Collision type and crash severity - 76 -

Figure 5-4: Number of crashes according to time of the day (*TIME*) - 77 -

Figure 5-5: Number of crashes during the week-end - 78 -

Figure 5-6: Seasonal distribution of crashes (*SEASON*) - 79 -

Figure 5-7: Number of crashes according to built-up area (*BUA*) - 80 -

Figure 5-8: Evolution of BIC, AIC and CAIC with the addition of clusters (Intersec_TL) - 83 -

Figure 5-9: Evolution of BIC, AIC and CAIC with the addition of clusters (Intersec_RS) - 85 -

Figure 5-10: Evolution of BIC, AIC and CAIC with the addition of clusters (Intersec_TR) - 87 -

Chapter 1: INTRODUCTION

The growth in road transport worldwide has brought about some negative impacts especially road traffic crashes. Despite measures implemented to reverse this situation, the external costs incurred still remain unacceptably high. Today, road crashes are a leading cause of premature deaths worldwide. The collection of relevant data regarding road crashes is a condition sine qua non for effective road safety strategies. Due to the complexity of road crashes, the data collected is usually very heterogeneous. Therefore, it becomes primordial to use statistical analytical techniques in order to retrieve useful information from this usually non-homogeneous traffic data. Moreover, certain sections of the road such as intersections or junctions are very important regarding the safety and mobility performance of a road network.

1.1 Background

Road transportation in Belgium facilitates the accessibility to jobs, education, markets, leisure and a wide range of other facilities thereby enhancing economic development. The advent of new Information and Communication Technologies (ICTs) and recent breakthroughs has further increased the role of transport as a key determinant of globalization. That notwithstanding, improvement in transportation systems has brought about other costs including crashes (accidents), pollution, noise, and environmental degradation.

Road traffic crashes still remain a major public health problem in Belgium despite a decline in fatalities per million population from 145 in 2001 to 90 in 2009 (ETSC, 2010). This trend is similar to several countries in Europe. In 2009, about 35 000 persons were killed in road crashes in the European Union (EU), while another 1.7 million people are recorded as injured on police records annually, with 300,000 being serious cases (ETSC, 2010). The estimated yearly cost of road traffic crashes to the EU member states is quite colossal and stands at more than € 180 million, which is roughly an equivalence of 2% of the Gross Domestic Product (GDP) of the EU (WHO, 2004). Due to these ravages, the EU in 2001 had set itself the ambitious target of reducing fatalities from crashes by 50% (27 000 cases) by the year 2010. Even though fatalities fell considerably by 15 400 cases in 2009 compared to 2001 (see figure 1-1); the

27 000 deaths limit set by the EU in its 2010 Road Safety Target was not attained. The annual progress since 2001 had been 4.4% implying that the target could be reached only until 2017 ceteris paribus (ETSC, 2009).

In a bid to create awareness and to reduce road crashes worldwide by half, the United Nations declared 2011-2020 as the Decade of Action for Road Safety. In line with this, the EU Road Safety Action Programme 2011-2020 includes challenging targets for the reduction of serious injuries and fatalities. In this regard, member states are encouraged to formulate national road safety strategies which are in line with the goals of the EU through the exchange and dissemination of good practices.

This research is centered on the application of a cluster analytical technique on crash data. Latent Class Cluster Analysis (LCCA) is used to model crashes that occurred in Belgium, specifically on road intersections using data for the year 2005. The main purpose is to identify the different crash or accident types that are dominant at specific types of intersections or junctions.

1.2 General Aspects of Road Safety

There are certain peculiarities of the general road safety situation which are a major cause for concern. For instance, the majority of road accident victims are young drivers below the age of thirty. A joint report from the Organization of Economic Co-operation and Development and the International Transport Forum (OECD/ITF, 2006) states that annually, more than 8 500 young car drivers die in thirty OECD member states. The same report further states that, road crashes are the single major killer of 15-24 year olds in industrial countries. Moreover, these young drivers are not just a danger to themselves, but they also pose greater risk to their passengers and other road users. In the year 2002, 5% of road fatalities in Europe involved children under 15 years old thereby making road accidents the leading cause of death for children (ETSC, 2007). This implies that if appropriate measures are not adopted in time, a devastating blow will be dealt on the youthful, working population of many developed countries.

Pedestrians, bicyclists and motorcyclists are highly vulnerable to traffic accidents. Klop and Khattak (1999) and Räsänen and Summala (1998) stated that a high proportion of serious injury bicyclist crashes involve collisions with vehicles at intersections. Furthermore, old people are largely involved in accidents at intersection due to their

high susceptibility to injury, especially as pedestrians and cyclists (Oxley et. al., 2004). In a related study, Harkey (1995) pointed out that 33% of older pedestrian deaths and 51% of older pedestrian injuries occurred at intersections. As road users advance in age, understanding complex traffic situations such as that encountered at an intersection become problematic. Considering the fact that the population of most European countries is rapidly ageing as a result of longer life spans, this will pose a serious problem in future if appropriate measures are not taken in time. Significant improvement in road safety can only be realized if the plight of these vulnerable road users is ameliorated.

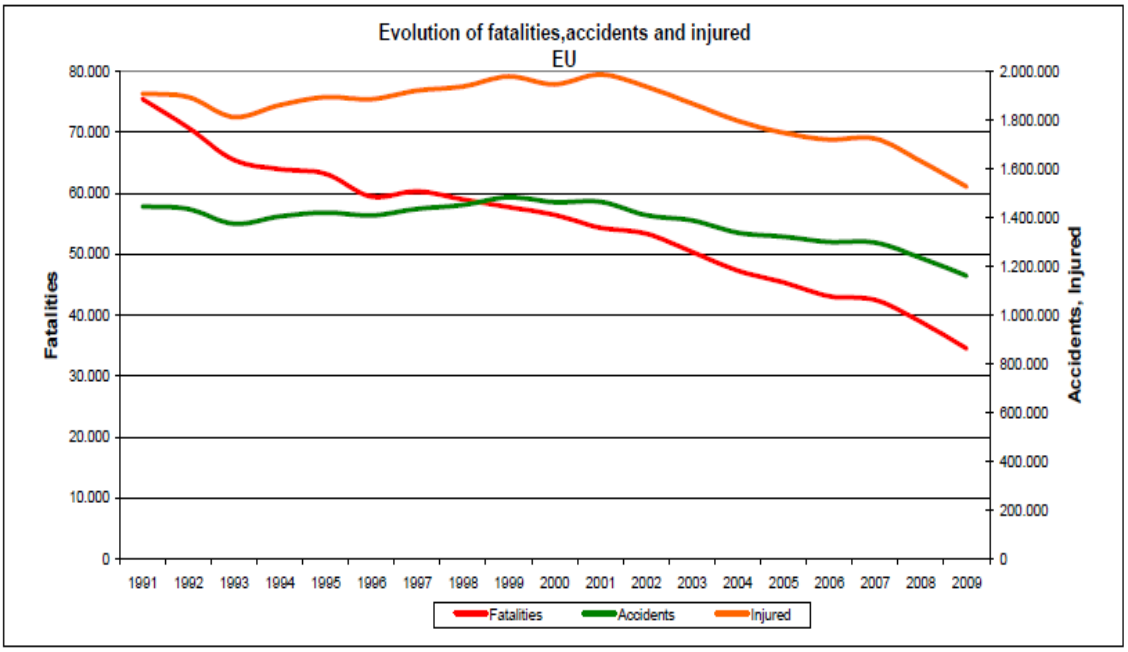

Figure 1-1: Evolution of road safety in Europe (1991-2009)

Source: CARE (EU road accidents database) or national publications
European Commission / Directorate General Mobility and Transport

Moreover, rapid urbanization and industrialization have exacerbated the existing road safety situation. The traffic safety problem is further aggravated by the rapidly increasing level of motorization especially private car use in Europe. Despite measures

implemented to discourage private car use, its ownership and use still remains high. For instance, in Europe, private car use has been expanding very rapidly (figure 1-2) in comparison to other modes of transport such as railway, air and sea.

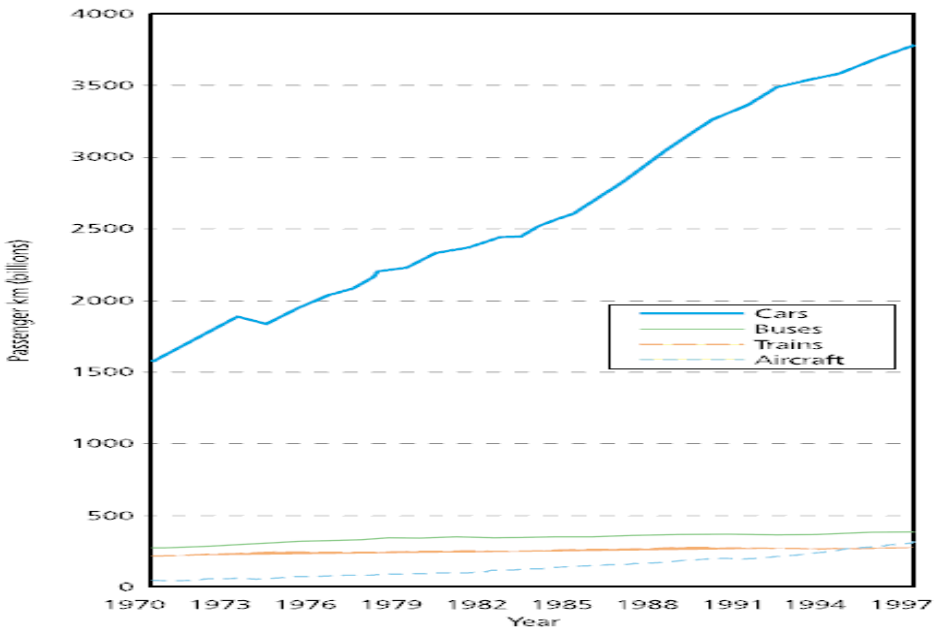

Figure 1-2: Increasing car use in Europe compared to other modes.

Source: Europe's environment: The second assessment (1998).

1.3 Intersections and Traffic Safety

This section deals with the role of intersections on traffic safety and the different types of intersections.

1.3.1 Road Safety at Intersections

Intersections or junctions are known to be one of the most dangerous locations on the entire road network. Due to the high frequency and severity of traffic collisions at intersections, they have been major targets of traffic safety strategies. Between 1996 and 2004, almost 61 000 persons were killed in traffic crashes at intersections in fourteen European countries, representing 21% of all the overall traffic accident

fatalities in these countries (ERSO, 2006). According to the same report, more than one-third of crashes (35.3%) registered in the United Kingdom in 2004 occurred at intersections.

Due to the merging of opposing traffic streams and the subsequent high number of conflict points, intersections are major causal points of various types of collisions including head on, rear-end and side collisions. As depicted on figure 1-3 below, conflict occurs where one vehicle path crosses, merges or diverges with, or queues behind the path of another vehicle, pedestrian, or bicycle. Therefore, intersections are the greatest point of conflict in traffic involving various types of road users. Depending on the type of intersection; signal controlled, unsignalised or round-about, the type and severity of the crash is likely to vary.

Round-abouts are generally better in terms of road safety than other intersection types. Cost-benefit studies carried out in several countries on the conversion of intersections into round-abouts have shown a general increase in road safety (Garder, 1998; Schoon and Van Minnen, 1993; Maycock and Hall, 1984). Elvik (2003) conducted a meta-analysis of the safety impacts of converting intersections to round-abouts outside the United States of America and concluded that the number of injury accidents reduced by 30-50% while fatal crashes considerably decreased by 50-70%. This is partly due to the reduction of conflict points on the road network. However, despite the reported increase in overall road safety, the safety level of certain road users has instead shown a decline. For instance, a before-and-after study carried out by Daniels, Nuyts and Wets (2008), on the safety impacts of the conversion of 95 intersections into round-abouts in Flanders (Belgium) showed that, even though the overall safety level improved, injury accidents involving bicyclists instead increased by 29%.

These studies clearly demonstrate the significance of intersections on road networks. Furthermore, the type of intersection that is constructed should reflect the main goals that the project seeks to achieve at the level of the target group of road users. However, technical impracticability can restrict the construction of certain types of intersections at specific locations.

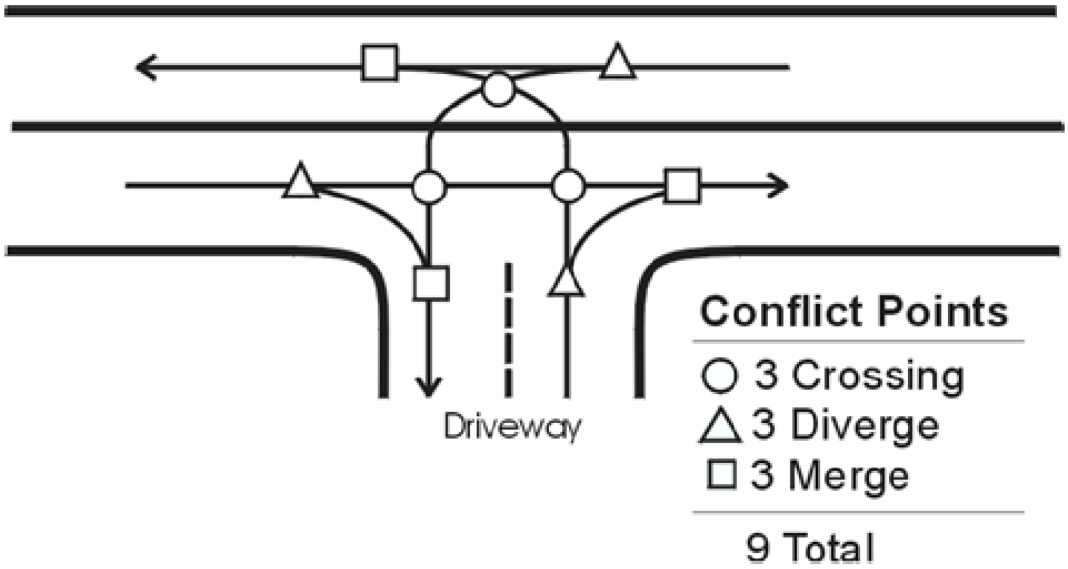

Figure 1-3: A T-intersection showing the different conflict points.

1.3.2 Types of Intersections

Intersections or road junctions have been classified in several ways. A common classification is based on the number of road segments or arms that meet at the intersection while others are based on the shape or the manner of traffic control. Generally, five types of intersections can be identified based on the type of traffic control (Bird, 2001).

- **Uncontrolled intersections**: These are junctions with no road signs or markings to indicate priority. This is usually common where the traffic volume is relatively low. However, in most cases, vehicles coming from the right have priority of way over the other vehicles.
- **Priority intersections**: Such intersections do have clear road signs and markings that indicate which vehicles are supposed to stop (B5) or give way (B1) to oncoming traffic. Hence, unlike uncontrolled intersections, roads with priority are clearly indicated.
- **Round-abouts**: These are circular junctions that direct traffic around a central island with several exit points. They vary in size depending on the flow of traffic and the number of road segments.

- **Signal-controlled intersections**: Intersections where the different traffic streams alternately have priority based on the changing signals. The use of traffic lights at intersections is very common especially on locations where the traffic volume is high.
- **Grade separated Junctions**: These are junctions where opposing traffic streams are separated by being at different levels. This can be accomplished by means of a bridge, tunnel or flyover.

  On the basis of shape, intersections can be classified into T or Y junctions, staggered junction, acute angle junction or multiple junctions among others..

1.4 Road Safety and Infrastructure in Belgium

1.4.1 Overview of Road Safety

Report from the Belgian Institute for Traffic Safety in 2000 cited in Geurts, Wets, Brijs and Vanhoof (2003) indicates that approximately 50 000 injury accidents occur annually in Belgium, involving some 70 000 victims with 1 500 deaths (Table 1-1). This is approximately 150 deaths per million population. However, road safety in Belgium has witnessed significant improvements over the years in line with the general trend in Europe. The highest number of traffic accident deaths was recorded in 1973 with an overwhelming 1 866 fatalities which dramatically dropped below 1000 for the first time in 2008 with 944 registered cases. (FPS Economy-Statistics Belgium, 2009).

Even though the vehicle population in Belgium has been on the rise (Table 1-3), the number of accidents has been relatively stable for many years. Moreover, the severity of accidents (seriously injured and deaths) has dropped significantly. A consistent decrease in fatalities was only realized after 2001 (Figure 1-4). Before this period, accident fatalities fluctuated from year to year. Between 2000 and 2009, road fatalities in this country decreased by 36%. Hence, fatalities per million population dropped from 145 in 2001 to 90 in 2009 (ETSC, 2010). Despite this progress, the number of those killed in 2009 is still 20% higher than the average of EU27. However, between 2000 and 2007, the number of deaths per million inhabitants decreased faster than the European average; -30% against -26% (FPS Economy-Statistics Belgium, 2008)

Table 1-1: Traffic accidents and casualties in Belgium

	2000	2005	2006	2007	2008
Number of injury accidents	49 064	49 313	49 181	49 815	48 827
Number of casualties	69 430	66 476	66 344	66 915	65 381
Dead*	1 470	1 089	1 075	1 071	944
Seriously injured	9 846	7 272	6 999	6 997	6 782
Slightly injured	57 588	58 114	58 270	58 847	57 654
Number of deaths according to road user					
Users of cars	922	624	594	550	479
Users of motorcycles	118	125	131	139	108
Users of mopeds	66	30	36	36	42
Cyclists	134	71	91	90	86
Pedestrians	142	108	123	104	99

* People killed within 30 days from the day of the accident.

Source: FPS Economy-Statistics Belgium, 2009.

The cost arising from accidents in Belgium in terms of lives lost and other costs such as property damage, medical and administration costs is consequently high, not to mention the psychological pain, grief and suffering it brings to victims and family members. According to Hoornaert (2010) even though the social costs of road crashes decreased by 14% between 2000 and 2008, that for vulnerable road users including pedestrians, bicyclists moped riders and motorcyclists instead witnessed an increase. The Belgian government through the regional authorities is therefore highly committed to improving the level of road safety especially for these vulnerable road users.

The causes of these crashes are wide ranging and principally involve human factors, vehicle and the road infrastructure. Excessive or inappropriate speeding is estimated to be a major cause of 30% of all road fatalities in Belgium (ETSC, 2008). Other related human factors are drink-driving, traffic light violations and multitasking while driving. Measures taken so far amongst others include reducing the speed limits especially around schools where it is currently 30km/h. Moreover, speed cameras have been in

use since 2001 to ensure compliance with speed limits and mitigation of other traffic violations. Above all, Intelligent Transport Systems (ITS) are currently used to reduce the possibility of accidents as well as reducing its severity by enhancing the protection of driver, occupants as well as other third parties when they actually occur.

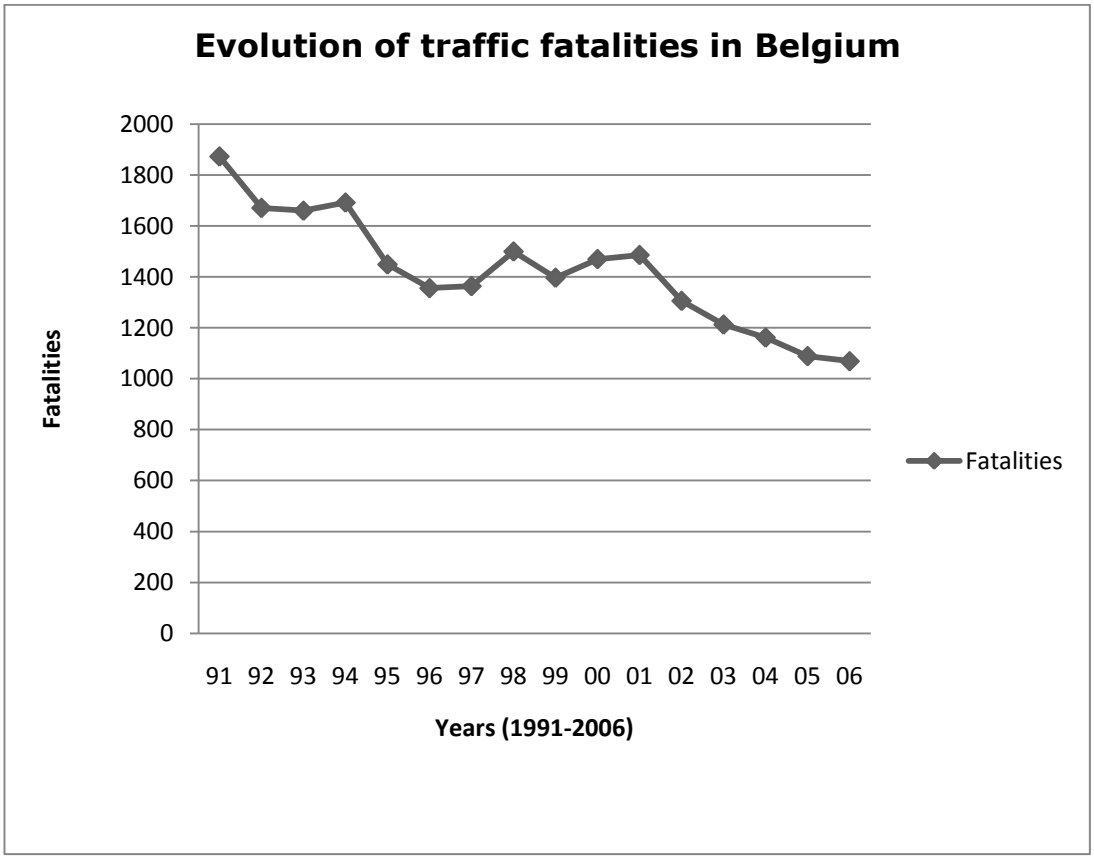

Figure 1-4: Evolution of traffic fatalities in Belgium (1991 to 2006).

Source: CARE Database

1.4.2 The Road Network

The Belgian road network is well developed, very dense and highly interconnected to other transport modes. Development of the road transport system has grown tremendously especially since after the Second World War when considerable trafficable road was damaged. Today, the transport infrastructure is well maintained with frequent

maintenance and reconstruction works and the roads are well lit. In 2005, the entire road network totaled 118 414 km (FPS Economy-Statistics Belgium, 2009).

Table 1-2: The Road network in Belgium (2005)

Road type	Length(km)
Motorways	1 747
Highways	13 892
Communal or street roads	102 775
Total	118 414

Source: www.statbel.fgov.be

Belgian roads are numbered as R-roads, A-roads, B-roads, N-roads and secondary N-roads or provincial roads. R-roads are ring roads around major cities and the Brussels ring denoted by RO is the most popular. It is a circular highway that surrounds the city of Brussels with other smaller towns on its southern flank. The ring stretches about 75km with two to three lanes in each direction and traverses the three regions of Brussels, Flanders and Wallonia. (http://en.wikipedia.org/wiki/Brussels_Ring).

A-roads are motorways which connect major cities and international destinations. They are not usually built as limited access facilities and may include traffic lights and grade crossings. A-road numbers radiate out from Brussels in a clockwise motion while those numbered from ten upwards radiates out from Antwerp in a similar manner. Belgium has a very dense network of motorway which is second only to the Netherlands (www.belgianroads.tk).

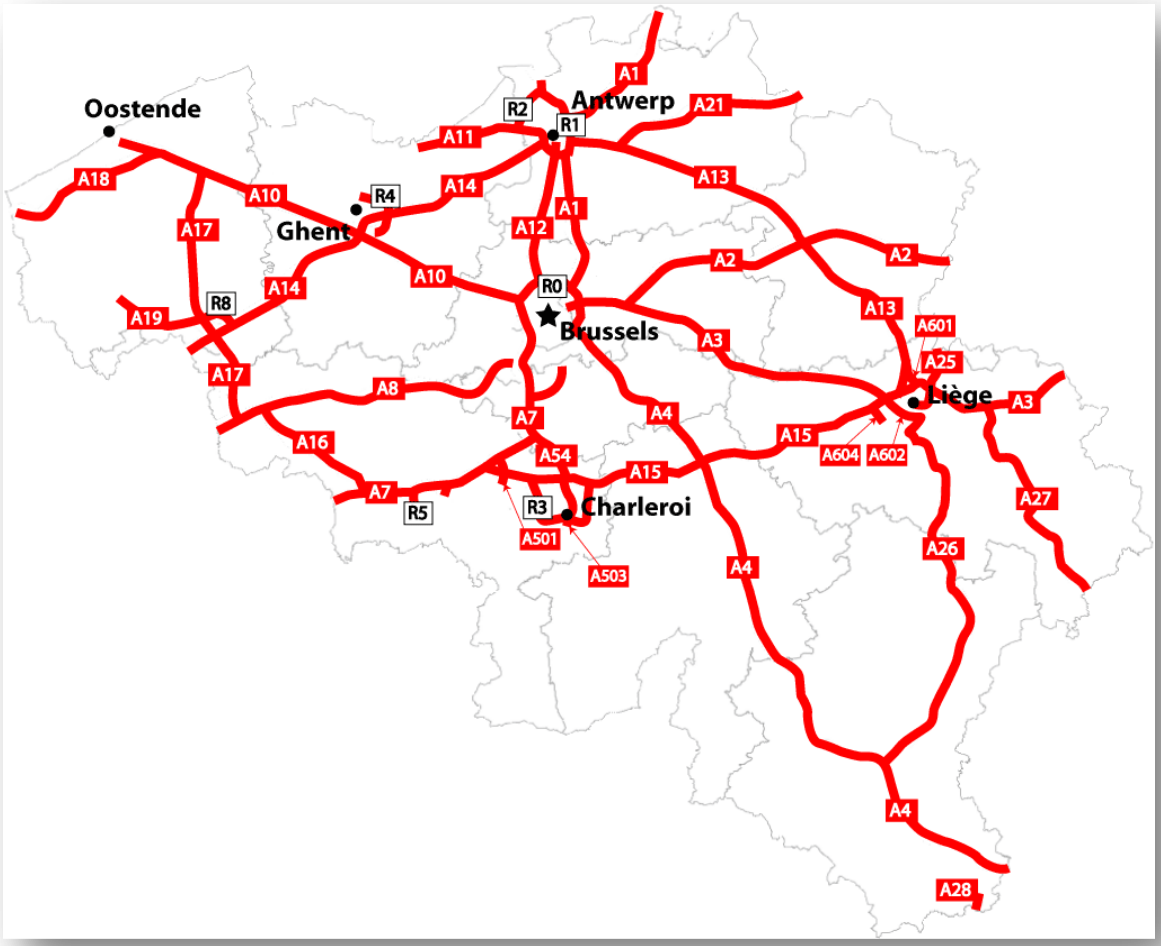

Figure 1-5: A-road system with major R-roads in Belgium

Source: www.europe.aaroads.com

B-roads are usually of expressway quality and act as short link roads between other points. N-roads can have motorway characteristics and grade separated interchanges but are mostly two lane roads connecting secondary cities and towns. N-roads approximately form a web and many converge on Brussels. Provincial routes are rarely sign posted, connecting small towns and villages and numbered according to which N-road they are nearest to. Most of the roads are connected with other European countries like France, the Netherlands, Germany and Luxemburg which constitute part of the Trans-European Transport Networks. Belgium highly utilizes the European

numbering system based on the International E-road Network with well designated roads. These roads are recognizable by the letter 'E' before the road number.

The rapid expansion of the road network has also been marked by an increase in car ownership especially in private cars. In spite of the crisis that rocked the car industry in Europe and other parts of the world in 2009, the car population in Belgium instead witnessed an increase. In 1977, there was one car per 3.5 inhabitants but by 2009, the ratio has increased to one car per two inhabitants. By mid 2009, about 6.5 million vehicles were plying Belgian roads of which more than 5 million were passenger cars as depicted on Table 1-3 below. (FPS Economy-Statistics Belgium, 2009). Therefore, the negative impacts arising from road transport poses a major problem. Moreover, it should be noted that the management of the Belgian road network is at the level of regional authorities.

Table 1-3: Motorization level in Belgium

On 1 August + evolution	**1977**	**1987**	**2000**	**2005**	**2006**	**2007**	**2008**	**2009**	**Growth 2009/2008**
Motor-vehicle population on 1 Aug.(inc. motorcycles).	3 315 071	4 158 127	5 735 034	6 158 742	6 251 428	6 362 161	6 482 033	6 574 789	1.4%
Passenger cars.	2 773 344	3 497 818	4 678 376	4 918 544	4 976 286	5 048 723	5 130 578	5 192 566	1.2%
Buses & coaches.	19 517	15 060	14 722	15 391	15 329	15 479	15 992	16 061	0.4%
Motor vehicles for goods transport (a).	236 421	556 397	502 979	604 437	623 250	642 687	662 780	676 644	2.1%
Tractors (b).	34 682	47 102	45 452	47 646	47 164	48 060	49 109	47 418	-3.4%
Agricultural tractors.	114 517	164 090	162 123	168 284	170 613	172 818	174 709	176 522	1.0%
Special motor vehicles (d).	32 489	57 432	53 544	58 147	59 022	59 651	60 585	61 638	1.7%
Motorcycles (c).	104 101	319 480	277 838	346 293	359 764	374 743	388 280	403 940	4.0%
Inhabitants per passenger car on 1 August.	3.55	2.15	2.19	2.12	2.11	2.10	2.08	-	-

(a) Trucks, vans, all-terrain vehicles, tank trucks.

(b) Road tractors are commercial motor vehicles to which semi-trailers (vehicle without front axle) are hitched.

(c) All motorcycles doing more than 40 km/h, i.e. all motorcycles and most mopeds.

(d) The special vehicles are slow vehicles the dimensions or weight of which exceed the normally allowed maximum values to transport goods.

It should be known that the maximum permissible weight in Belgium should not exceed 44 tonnes. The vehicles of this category are permitted to drive on the public highway only under very strict condition.

Source: FPS Economy – Statistics Belgium, 2009.

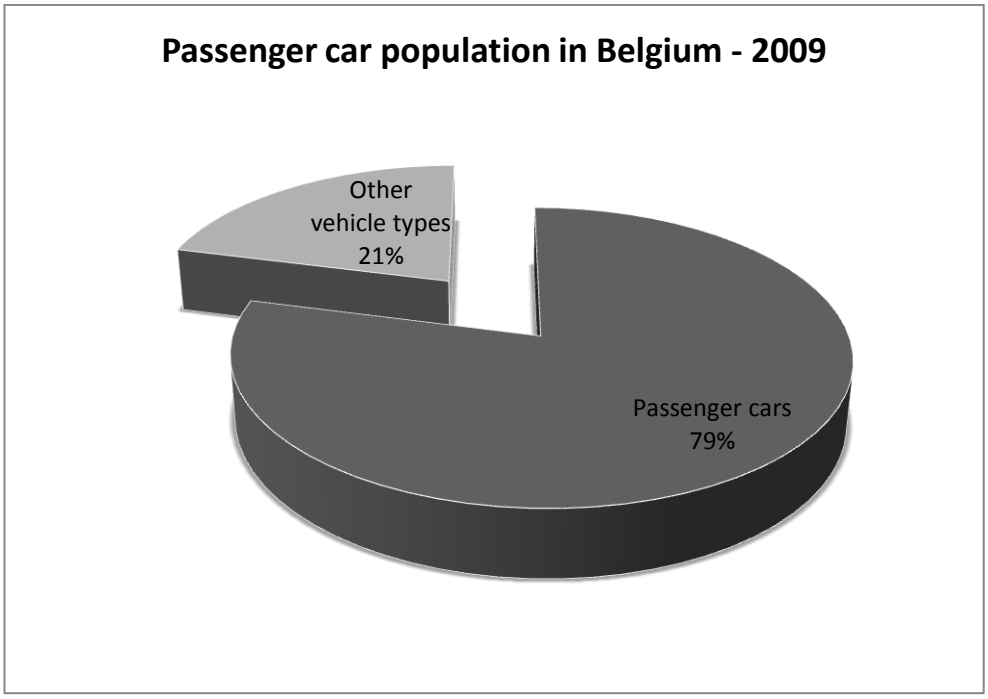

Figure 1-6: Passenger car population in Belgium - 2009

Source: FPS Economy – Statistics Belgium, 2009.

1.5 Scope of Study and Research Objectives

Current road safety figures clearly indicate that road crashes are on a decline in many western countries. That notwithstanding, there is still a need to further reduce these fatalities to the lowest rate possible. This calls for more robust and rigorous studies of those aspects that directly or indirectly impact on the occurrence and severity of road crashes. These factors are linked to the vehicle, road infrastructure, the behavior of road users and also to the environment. These factors do not act in isolation, but interact with each other to contribute to road crashes.

This study focuses on crashes that occurred at intersections. Intersections play an important role in the mobility and safety performance of a road network depending on

the particular type. Some studies on road safety (Elvik, Hoye, Vaa and Sorensen, 2009, Kulmala, 1994) assert the need for redesigning some intersections as a road safety strategy. Moreover, knowing the specific characteristics of crash types that occur at the different types of intersection is very valuable when undertaking such safety measures as they will likely differ in their degree of injury severity. For instance, some accident types might show a high proportion of head-on collision, which is known to be more fatal than side-collision or rear-end collision. Hence, strategies can be aimed at reducing the occurrence of certain specific types of collisions. Another scenario might be an intersection where human errors are frequently mentioned to be partly the cause of crashes. In this case, making the intersection more road user friendly can be an option for road safety authorities.

This research is based on crash data for Belgium and seeks to achieve the following objectives:

- To demonstrate the usefulness of Latent Class Cluster Analysis (LCCA) as an effective statistical method to cluster or segment accident data.

- To identify the dominant crash or accident types that are associated with particular types of Traffic control at intersections.

- To describe the accident data of crashes that occurred at intersections.

- Finally, to make a comparative assessment of dominant crash types among different types of intersections.

  In a nutshell, this model-based clustering technique will be used to classify a sample crash data on intersections in Belgium into homogenous groups. From this, the different accident types based on traffic control at the intersections will be deduced. To this end, informed decision can be made regarding the safety potential of intersections in particular, and the entire road network in general.

1.6 Research Method

Cluster analysis is a useful analytical tool for discovering structure or pattern within large heterogeneous data sets. It involves classifying objects or observations into homogenous groups based on certain criteria. This research utilizes a novel technique of cluster analysis known as Latent class cluster analysis (Vermunt and Magidson,

2002). This technique is also known by different appellations notably finite mixture models (Fraley and Raftery, 2002), mixture densities-based clustering (Xu and Wunsch, 2005), model-based clustering (Banfield and Raftery, 1993) and Bayesian Classification (Cheeseman and Stutz, 1995).

Latent class cluster analysis (LCCA) is a model-based and more efficient clustering technique compared to traditional distance-based clustering methods such as the K-means and hierarchical clustering. In LCCA, an observation's posterior class membership probabilities are computed from the estimated model parameters and its observed scores (Vermunt and Magidson, 2002). Therefore, observations belong to several clusters with varying degree of membership and are assigned to the cluster with the highest membership probabilities.

This technique is used to model crashes that occurred at intersections in Belgium in the year 2005. After selecting variables for the model, the data set is then split up according to the different intersection types. The next step involves identifying the crash types that are common to specific types of intersections. However, a pre-analysis using descriptive statistics including frequency tables and charts is first conducted to obtain some important notions about the data. The data is then processed using the statistical software package known as Latent Gold (Vermunt and Magidson, 2000). Latent Gold processes data by running several iterations until convergence is attained. The model selection is based on low BIC, AIC and CAIC values. The clusters obtained at the end of the process conceptually refer to different crash types.

1.7 Structure of the Report

After the background information provided in this chapter, the rest of the report is structured as follows:

Chapter two examines the traditional or distance based clustering techniques. Some applications in the field of transport are mentioned, narrowing down to more specific applications in the domain of traffic safety. This is followed by a discussion of the various types as well as their strengths and weaknesses.

Latent class model (LCM) and its application in clustering; Latent class cluster analysis (LCCA) is covered in chapter three. The procedure of LCCA is fully examined.

The methodological aspect of this report is the focus of chapter four. The data sources, preparation, variable selection and processing is examined. Moreover, the advantages

of LCCA over traditional clustering techniques are highlighted. Next a description is given about clustering using the software package, Latent Gold (Vermunt and Magidson, 2000) and how the data was actually processed.

Chapter five presents the results of the analysis beginning with descriptive statistics, followed by results from Latent Gold. Next is a discussion of the results.

Finally, chapter six gives a conclusion including major findings of the research and some directions for further research.

Chapter 2: STANDARD CLUSTER ANALYSIS

Cluster analysis or clustering embodies diverse techniques for discovering structure or pattern within complex bodies of data. As stated in Xu and Wunsch (2005), cluster analysis has been applied in several fields including life and medical sciences (genetics, biology, microbiology, paleontology, clinic, pathology), computer sciences (web mining, spatial database analysis, image segmentation), engineering (machine learning, artificial intelligence, pattern recognition, mechanical and electrical engineering), earth sciences (geography, geology, remote sensing), social sciences (education, sociology, psychology, archeology) and economics (business, studies), (Moustaki and Papageorgiou, 2005; Green, 2004; Jiang, Tang and Zhang, 2004; Everitt, Landau and Leese, 2001; Arabie and Hubert, 1994; Hartigan, 1975.) For instance, Cluster analysis is used in Marketing to identify persons with similar purchasing needs. Based on this information, it becomes relatively easy to formulate efficient marketing strategies in the future. In Traffic Safety, Cluster analysis can be used to classify road crashes by age, gender, road user or severity level. From this, specific safety strategies or countermeasures can be formulated for different target groups.

Standard clustering techniques have also been applied in the field of transport. For instance, Esnaf, Koldemir, Küçükdeniz and Akten (2008) examined shipping accidents in relation to the different types and locations where accidents occurred frequently by applying a fuzzy clustering technique. Based on their analysis, accident characteristics at different locations were studied and key factors behind sea accidents in the Bosporous were obtained.

Weijermars and Van Berkum (2005) used the hierarchical clustering method to analyze highway flow patterns in the Netherlands. By collecting data on speed and flow on a daily 15 minutes interval, they obtained three clusters with distinct traffic profiles. In a related study, Chengqian, Jingling, Zhong and Yue (2009) in their research studied the traffic flow characteristics of a city tunnel by applying cluster analysis and other data mining techniques.

In the domain of traffic safety, Golob and Recker (2003) made use of clustering methods alongside Principal Component Analysis (PCA) and Non-linear canonical analysis (NLCA) to find relevant patterns in the relationship between accident type and

traffic flow characteristics on urban freeways. In a similar study, Postorino and Sarnè (2001) analyzed different accident types using clustering techniques to decipher the principal causes of accidents, by classifying accidents into homogenous groups. In a more recent study, Kleinemas and Rudinger (2010) applied clustering methods to profile elderly drivers' involvement in road accidents. From this, they were able to get an insight into the behavior of old drivers.

Several studies have been conducted to analyze road crashes using statistical models (Savolainen, Mannering, Lord and Qudus, 2011). For example, Kim, Nitz, Richardson and Li (1995) formulated a log-linear model to identify the relationship between driver characteristics and behaviour and the severity of traffic crashes. Their results indicated that alcohol or drug use and non seat belt use sharply increases the odds of more severe crashes and injuries. In another study, Roh, Bessler and Gilbert (1999) demonstrated the benefits of using statistical methods based on directed graphs to model traffic accident fatalities.

Ossenbruggen, Pendharkar and Ivan (2001) in performing a risk assessment of a region, made use of a logistic regression model to identify factors that were statistically significant in predicting the probability of crashes and injury. Using factors like land use activity, roadside design, use of traffic control devices and traffic exposure; they concluded that village sites are less hazardous than residential and shopping streets.

In a related study, Bédard, Guyatt, Stones and Hirdes (2002) used a Multivariate Logistic regression model to identify the separate contribution of driver, crash and vehicle characteristics to the fatality risk of drivers. They found out that consistent seat belt use, speed reduction and reduction of frequency and severity of driver side impacts will greatly minimize driver fatalities. Furthermore, to analyze the association between fatal crash rate (fatal crash per vehicle mile traveled) and speed limit in the state of Washington D.C, Ossiander and Cummings (2002) applied a Poisson regression technique. Their result confirmed that speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in that state.

However, due to the complexity of road crashes, some classic statistical models cannot properly be used to model crash data due to the existence of so many variables. As Chen and Jovanis (2002) showed, problems may arise when classic statistical analysis is used on data sets with many dimensions. They cited the exponential increase in the number of parameters with the addition of variables and the invalidity of statistical tests when there is sparse data in large contingency tables. Moreover, commonly used

statistical methods including Logistic regression, Poisson regression and Binomial regression, usually rely on certain assumptions and pre-defined underlying relationships between dependent and independent variables. Therefore, these models could lead to erroneous estimation of the likelihood of accident occurrence under the hypothesized conditions when these assumptions are violated (Prato et al., 2010).

This chapter gives an overview of standard clustering methods. The next section presents some early definitions of cluster analysis followed by assessment of *distance* in section 2.2. Distance metrics for continuous and binary variables are the focus of section 2.3 and finally, the types of clustering algorithms; their computation, advantages and disadvantages are covered in the remaining sections.

2.1 Definition of Cluster Analysis

Cluster analysis (Tryon, 1939) is a family of methods for organizing data into structures that are hoped to be meaningful. This early definition has been improved and modified over the years, but the underlying concept has not altered significantly. Cluster analysis or clustering is an exploratory data analytical technique aimed at extracting hidden information out of large and usually multi-dimensional data sets. It is used to classify a set of items into two or more mutually exclusive unknown groups or *clusters* based on a combination of interval variables. Cluster analysis has also been defined as the classification of similar objects into groups, where the number of groups, as well as their form is unknown (Kaufman and Rousseeuw, 1990). The *form of a group* according to this early definition refers to the cluster parameter; that is, to its cluster-specific means, variances, and covariances that also have a geometrical interpretation. Therefore, a cluster is a collection of objects which are homogeneous between items of the group and heterogeneous to items of other groups.

Graepel (1998) stated a simple, formal, mathematical definition of clustering as follows: let $X \in R^{mxn}$ equal a set of data items representing a set of m points x_i in R^n. The goal is to partition X into K groups C_k such that every data belonging to the same group is more '*alike*' than data in different groups. Each of the K groups is called a *cluster*. The result of the algorithm is an injective mapping $X \rightarrow C$ of data items X_i to clusters C_k. The number K might be pre-assigned by the user or it can be an unknown, determined by the clustering algorithm. Generally, cluster analysis is made up of two major components, which are the similarity (or distance) measure and the clustering algorithm.

The main goal of cluster analysis is to organize items into groups in such a way that the degree of similarity is maximized for the items within a group and minimized for those between groups. Moreover, this technique is often used when the researcher is interested in the characteristics of the individual items in the dataset, rather than aiming to test causal hypotheses. Therefore, the aim is to summarize the data as simply, practically and effectively as possible, rather than to estimate particular quantities.
Due to the fact that the actual number of clusters as well as their form is unknown, it is a classic example of unsupervised learning (Vermunt and Magidson, 2002). Hence it is also referred to as unsupervised classification (Zaiane, 1999). Unsupervised learning which is a sub-field of machine learning, studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input pattern (Dayan, 1999). Contrary to supervised or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output.

Clustering is a core task in the data mining procedure for discovering groups (Han and Kamber, 2000). Data mining which in turn is a step of the Knowledge discovery in databases (KDD) process refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from a large database (Frawley, Pietetsky-Shapiro and Matheus, 1992). It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans. Data mining techniques are now widely used in a variety of fields including traffic safety as a result of the need for analyzing massive and multidimensional data and the development of fast computing algorithms. Knowledge discovery in databases simply refers to the process of finding useful information and patterns in data (Dunham, 2003).
Other data mining techniques that have been increasingly applied in the domain of traffic safety and mobility management include:

- Artificial Neural Networks (ANN) (Moghaddam, Afandizadeh and Ziyadi, 2011; Chimba and Sando, 2009; Delen, Sharda and Bessonovi, 2006; Bayam, Liebowitz and Agresti, 2005; Chang, 2005; Chong, Abraham and Parzycki, 2005; Mussone, Ferrari and Oneta, 1999).
- Classification and Regression Trees (CART) or Decision trees (Chang and Wang, 2006; Chang and Chen, 2005; Chong, Abraham and Parzycki, 2005;

Yamamoto, Kitamu and Fujii, 2002; Zeitouni and Chelghoum, 2001; Dougherty, 1995; Yang, Kitamura, Jovanis and Vaughn, 1993).

- Association rules (Kavsek, Lavrac and Jovanoski, 2006; Geurts, Wets, Brijs and Vanhoof, 2003a) Kohonen networks (Tseng, Nguyen, Liebowitz and Agresti, 2005), Frequent item sets (Geurts, Thomas and Wets, 2005).

In a nutshell, Cluster analysis is a technique used for grouping observations such that:

- Each group is homogeneous with respect to certain characteristics; that is observations in each group are similar to each other.
- Each group is different from other groups (heterogeneous) with respect to particular characteristics; that is observations of one group differ from those of other groups.

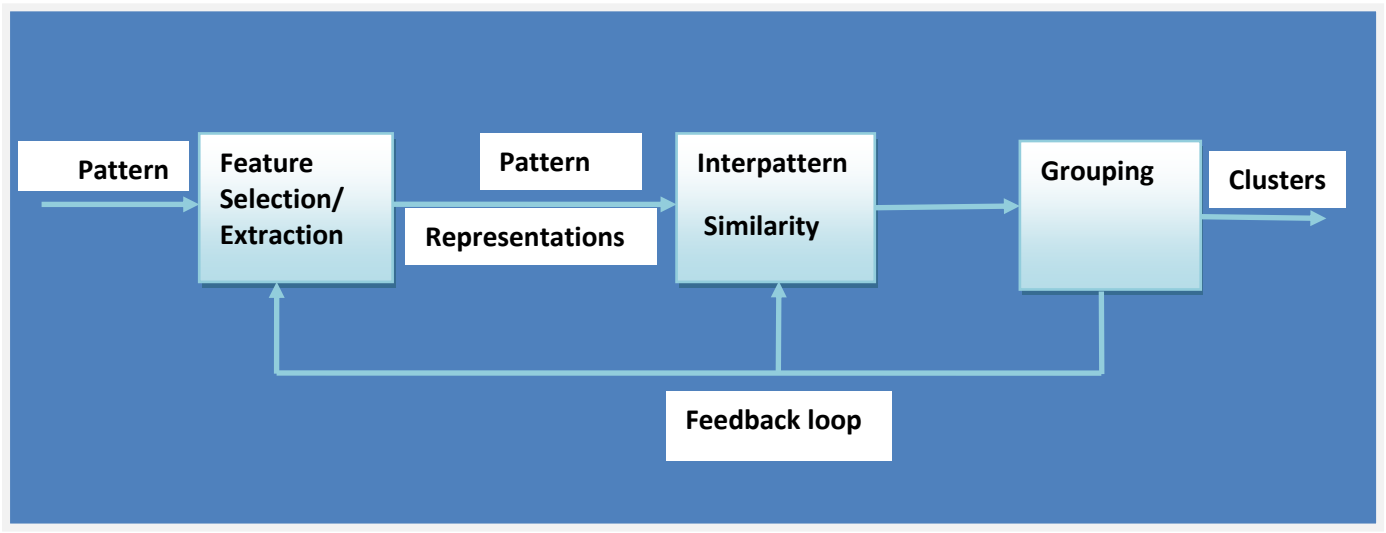

Figure 2-1: Stages in Clustering
Source: Jain, Murty and Flynn, 1999.

2.2 Assessing Similarity and Dissimilarity in Cluster Analysis

As earlier stated, cluster analysis seeks to maximize the similarity within clusters while minimizing that between clusters. Most often, the level of similarity is assessed by specifying a specific distance function when the variables in the study are continuous in nature and a similarity matrix when the variables portray qualitative characteristics. When both continuous and qualitative features are present, a mapping is applied on to the interval (0,1) such that a distance measure can be utilized.

The idea of dissimilarity captured by distance forms an essential component of clustering algorithms and permits one to navigate through the data space and form clusters. Using the dissimilarity measure, it is possible to sense and articulate the closeness of two patterns, and based on this allocate them to the same cluster. Hence, lower values indicate more similarity.

Formally, the dissimilarity between x and y, denoted by d(x, y) is considered to be a two-argument function, satisfying the following condition:

$$d(x, y) \geq 0 \quad \text{for every x and y} \tag{2.1}$$

$$d(x, x) = 0 \quad \text{for every x} \tag{2.2}$$

$$d(x, y) = d(y, x) \tag{2.3}$$

To summarize, the dissimilarity measure should be a non-negative character as demonstrated above in equation 2.1. Another obvious requirement is the symmetry. When dealing with two identical patterns, the dissimilarity attains a global minimum, that is d(x, x) = 0.

Metric distance is a more restrictive concept as the triangular inequality has to be satisfied. For instance, for any pattern x, y and z;

$$d(x, y) + d(y, z) \geq (d, z) \tag{2.4}$$

2.3 Selected Distance Functions between Patterns x and y

Several distance metrics are used when performing cluster analysis. The objective is to compute the similarity or dissimilarity between two or more points and group similar points into clusters. Different metrics are used for continuous and binary variables as elucidated in the proceeding section.

2.3.1 Distance metrics for continuous variables

Several distance metrics are used when performing cluster analysis for continuous variables. These include amongst others the Euclideaan distance, Squared Euclidean distance, Mahalanobis (Nadler and Smith, 1993), Minkowski (Batchelor, 1978), Tchebyshev, Canberra, Manhattan, Angular and Hamming distance metrics (Michalsky, Stepp and Diday, 1981; Diday, 1974).

- **Euclidean distance (L_2 Norm):** This refers to the straight line distance between two points and it is the most common distance metric. Euclidean distance or simply distance "*as the crow flies*" examines the root of squared differences between coordinates of a pair of objects. One weakness of the basic Euclidean distance function is that if one of the input attributes has a relatively large range, then it can overpower the other attributes. For example, if an application has just two attributes, A and B and A has values from 1 to 1000, and B has values only from 1 to 10, then B's influence on the distance function will usually be overpowered by A's influence. Therefore, distances are often normalized by dividing the distance for each attribute by the range (i.e., maximum-minimum) of that attribute, so that the distance for each attribute is in the approximate range 0-1. In order to avoid outliers, it is also common to divide by the standard deviation instead of range, or to "trim" the range by removing the highest and lowest few percent (e.g. 5%) of the data from consideration in defining the range (Atkeson, C.G., Moore, A.W. and Schaal, S., 1997). It is also possible to map any value outside this range to the minimum or maximum value to avoid normalized values outside the range 0-1. Domain knowledge can often be used to decide which method is most appropriate. Related to the idea of normalization is that of using attribute weights and other weighting schemes. Many learning systems that use distance functions incorporate various weighting schemes into their distance calculations (Wettscherek, Aha and Mohri, 1995).

  The formula for the Euclidean distance between points $x = (x_1, x_2, \ldots, \mathrm{n})$ and $= (y_1, y_2, \ldots, \mathrm{n})$ is given as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (2.5)$$

Where;

n =number of variables and

xi and yi are the values of the ith variable at points x and y respectively.

- **Squared Euclidean distance:** This merely involves squaring the simple Euclidean distance in order to place progressively greater weight on objects that are further apart. As a result, clustering with the Euclidean squared distance metric converges faster than clustering with the normal Euclidean distance. Moreover, the output for k-means clustering does not change when either metrics are used. However, the output for hierarchical clustering is likely to change. This distance is computed as:

$$d(x,y) = \sum_{i=1}^{n}(xi - yi)^2 \quad (2.6)$$

- **Manhattan or City-block distance (L_1 Norm):** The Manhattan distance metric between two items is simply the sum of their differences across dimensions. In other words, it computes the distance that would be travelled to get from one data point to the other if a grid-like path is followed. Assuming again $X=(x_1, x_{2,}\ldots\mathrm{n})$ and $Y=(y_1, y_2,\ldots\mathrm{n})$ are two points, the Manhattan distance metric is computed as:

$$d(x,y) = \sum_{i=1}^{n}|xi - yi| \quad (2.7)$$

Where;

n, xi and yi are defined as above (Equation 2.5)

Usually, it yields similar result with the simple Euclidean distance. However, the effect of outliers is dampened since they are not squared.

(a) Manhattan distance (b) Euclidean distance

Figure 2-2: Distinction between Manhattan and Euclidean distance metrics

- **Tchebyschev distance (L_{max} or L_∞ Norm):** This distance metric examines the absolute magnitude of the differences between coordinates of a pair of points in any single dimension. It is also referred to as the maximum value or chessboard distance and was named after the Russian Mathematician Pafnuty Tchebyshev. It can be used for both ordinal and quantitative variables and is computed as:

$$d(x,y) = \mathrm{Max}_{i=1,2,\ldots,n} |x_i - y_i| \quad (2.8)$$

The Tchebychev distance metric may be adequate when the difference between points is reflected more by differences in individual dimensions rather than all the dimensions considered. It is also very sensitive to outliers.

- **Minkowski distance:** The Minkowski distance (Kruskal, 1964) is a generalized metric that includes other distance measures as special cases of the generalized form. Theoretically, infinite measures exist by varying the order of the equation,

but only three are of major importance in practice. Even though the Minkowski metric can be defined for any λ>0, it is rarely used for values other than 1, 2 and ∞.

It is expressed as:

$$d(x,y) = \sqrt[\lambda]{\sum_{i=1}^{n}(xi - yi)^{\lambda}} \quad , \qquad \lambda > 0 \tag{2.9}$$

λ=the order of the distance metric

Therefore when:

λ=1, it equals the Manhattan distance.

λ=2, it equals the Euclidean distance.

λ=∞, it equals the Tchebyshev distance.

This distance metric is mostly used when variables are measured on ratio scales with an absolute zero value. The result is affected when variables have a wide range and as such, a few outliers with high values bias the result and disregard the alikeness given by a couple of variables with a lower upper bound.

- **Canberra distance:** The Canberra distance metric (Lance and Williams, 1967) examines the sum of series of a fraction *'differences'* between coordinates of a pair of points. Each term of the fraction difference has value between 0 and 1.However, the Canberra distance itself does not lie between 0 and 1. If one of the coordinates is zero, the term becomes unity regardless of the other value, thus the distance will not be affected. Moreover, if both coordinates are equal to zero, it is written as 0/0 =0 and not infinity. This distance metric is very sensitive to a small change when both coordinates approach zero. The formula is written as:

$$d(x,y) = \sum_{i=1}^{n} \frac{|xi - yi|}{|xi + yi|} \, , xi \; and \; yi \; are \; positive \tag{2.10}$$

- **Mahalanobis distance:**

The Mahalanobis distance is also a common generalized distance measure. Also known as quadratic distance, it measures the separation of two groups of objects by taking into account the covariance among the variables in computing the distances. Hence, the problem of correlation and scale which is inherent in Euclidean distance is eliminated. It was named after the Indian Scientist and Statistician P.C Mahalanobis in 1936. Suppose there are two groups with means $\bar{x}_i$ and $\bar{x}_j$, then the Mahanalobis distance is computed as:

$$dij = ((\bar{x}i_i - \bar{x}j)^T \ S^{-1} (\bar{x}i - \bar{x}j))^{1/2} \qquad (2.11)$$

where,

S = covariance matrix

By choosing this matrix, the geometry of potential clusters can be controlled by rotating the ellipsoid (off diagonal entries of S) and changing the length of its axes (the elements lying on the main diagonal of the matrix). When the covariance matrix is the identity matrix, the Mahanalobis distance is equal to the Euclidean distance.

The afore-mentioned distance functions (section 2.3.1) are useful when the variables concerned are continuous. Each of these functions implies a different view of the data because of their geometry. The geometry is easily illustrated when only two features are considered (For example $X = [x_1 x_2]^T$) and the distance of x from the origin is computed. The contours of the constant distance (figure 2-3) show what type of geometric construct becomes a focus of the search for structure. Here we become aware that the Euclidean distance favors circular shapes of data clusters. With the distance functions come some taxonomy or classification, for instance the Minkowski distance comprises an infinite family of distances, including well-known and commonly used ones such as the Hamming, Tchebyschev, and Euclidean distances.

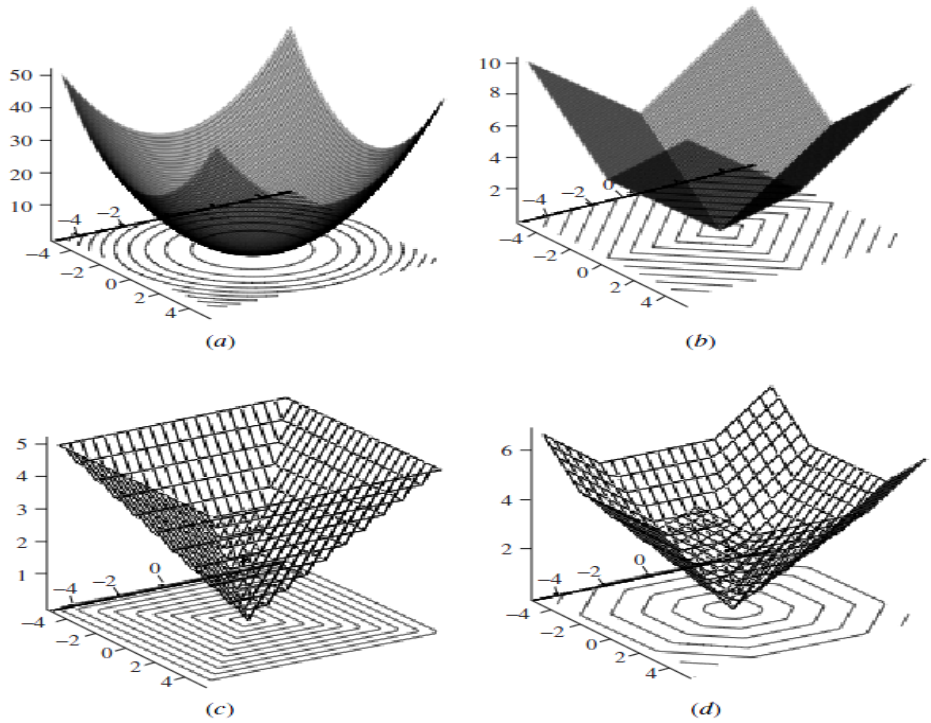

Figure 2-3: Examples of distance functions-three dimensional and contour plots: (a) Euclidean, (b) Hamming, (c) Tchebyschev (d) 'Combined' type of distance measure (2/3 Hamming, Tchebyschev

Source: Pedrycz, 2005.

2.3.2 Distance measures for binary variables

With binary variables, the general focus is on the notion of similarity rather than distance (or dissimilarity). Consider two binary vectors x and y that consist of two strings [xk], [yk] of binary data; compare them coordinate wise and do the simple counting of occurrences:

Number of occurrences when x_k and y_k are both equal to 1
Number of occurrences when $x_k = 0$ and $y_k = 1$
Number of occurrences when $x_k = 1$ and $y_k = 0$
Number of occurrences when x_k and y_k are both equal to 0

These four numbers can be organized in a 2 by 2 co-occurrence matrix (contingency table) that visualizes how *"close"* these two strings are to each other.

	1	0
1	a	b
0	c	d

Evidently, the zero non-diagonal entries of this matrix point at the ideal matching (the highest similarity). Based on these four entries, there are several commonly encountered measure of similarity of binary vectors x and y. These are :

- The Matching coefficient:

$$\frac{a+d}{a+b+c+d} \tag{2.12}$$

- The Russell and Rao measure of similarity consists of the quotient:

$$\frac{a}{a+b+c+d} \tag{2.13}$$

- The Jacard index involves the case when both inputs assume values equal to 1.

$$\frac{a}{a+b+c} \tag{2.14}$$

- The Czekanowski index is practically the same as the Jacard index, but by adding the weight factor of 2, it emphasizes the coincidence of situations where entries of x and y both assume values equal to 1:

$$\frac{2a}{2a+b+c} \tag{2.15}$$

2.4 Types of Clustering Algorithms

Cluster analysis involves many different techniques and the procedure will depend on the particular technique used. However, there are two major categories of clustering techniques. Hierarchical (nested) and Non-hierarchical (Partitional) clustering algorithms. With hierarchical clustering, successive clusters are found using previously established clusters whereas partitioning algorithms determine all clusters at once. Partitioning methods partition the data into pre-specified number of clusters (k) of mutually exclusive and exhaustive groups. On the other hand, hierarchical methods do not specify how many clusters are appropriate a priori, but clusters are obtained by "*cutting"* the tree at some level as shown on the dendogram below (figure 2-4).

2.4.1 Hierarchical clustering

Hierarchical cluster analysis (or nested clustering) is a general approach to cluster analysis, in which the objective is to group together objects or records that are *"close"* to one another. A key component of the analysis is repeated calculation of distance measures between objects, and between clusters once objects begin to be grouped into clusters. The clustering techniques in this category produce a graphic representation of data (Duda, Hart and Stork, 2001). The outcome is represented by a tree diagram known as a dendogram as shown on figure 2-4. A dendogram is a graphical representation of the results of a hierarchical cluster analysis. It is a tree-like plot where each step is represented as a fusion of two branches of the tree into a single one. The branches depict clusters obtained on each step of the hierarchical clustering algorithm.

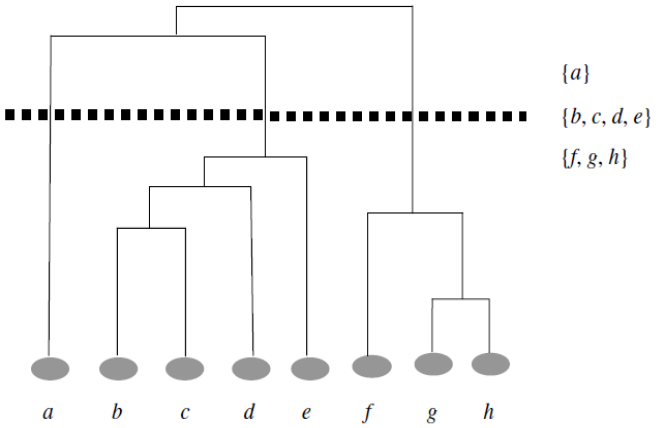

Figure 2-4: A Dendogram

The hierarchical clustering technique is further divided into two types:

2.4.1a Agglomerative or Bottom-up Technique:

It starts with one point or singleton cluster and recursively merging two or more most similar clusters to one *"parent"* cluster until the termination criterion is reached. For example, until k clusters are built. This method is often used is practice. Below is a simple agglomerative algorithm:

- Initialise the cluster set assuming that each point is a distinct cluster.
- Compute the similarity between all pairs of clusters that is, calculating the similarity matrix whose ij^{th} entry gives the similarity between the i^{th} and j^{th} clusters.
- Merge the most similar (closest) clusters.
- Update the similarity matrix to reflect the pair-wise similarity between the new cluster and the original (remaining) clusters.
- Repeat steps 3 and 4 until only a single cluster remains.

2.4.1b Divisive or Top-down Technique:

Contrary to the agglomerative approach, it starts with one cluster of all objects and successively splitting each branch until the termination criterion is attained. This method is not frequently used in practice.

The initial data for the hierarchical cluster analysis of N objects is a set of ***N x N (N – 1) /2*** object-to-object distances and a linkage function for computation of the cluster-

to-cluster distances. The linkage function is an essential pre-requesite for hierarchical cluster analysis. Its value is a measure of the *"distance"* between two groups of objects or clusters. Three are five major types of linkage function used in hierarchical clustering.

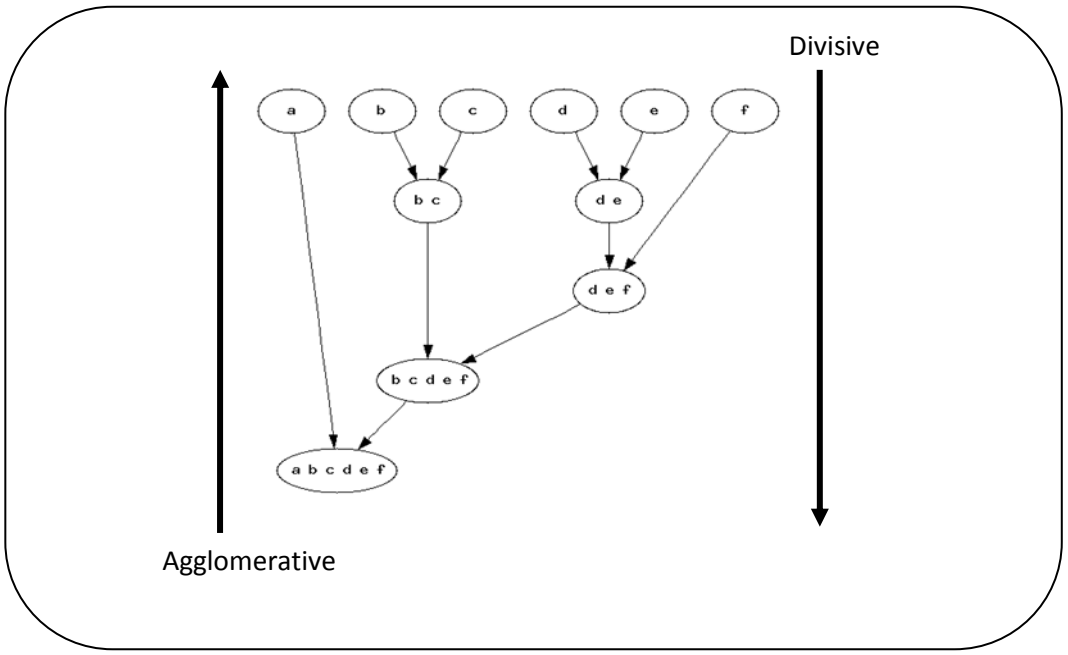

Figure 2-5: Hierachical Clustering

2.5 Types of Linkage Functions

In hierarchical clustering, the decision to merge or split clusters depends on the linkage function used. Several of these linkage functions are available and include Single-linkage, Complete linkage, Average linkage, Group average linkage and Ward's method amongst others.

- **Single Linkage or Nearest-Neighbor Method (MIN):**

Based on this method, the dissimilarity between two clusters is the minimum distance between objects of the two clusters. The distance is computed as the minimal object-to-object distance, $\boldsymbol{d(x_i, y_j)}$ where objects $\mathbf{x_i}$ belong to the first cluster, and objects $\mathbf{y_i}$ belong to the second cluster. In other words, this technique computes the maximum similarity between two groups of objects or clusters. This method produces long chains which form loose, straggly clusters. This technique is useful in handling non-elliptical shapes, but it is very sensitive to noise and outliers. Clustering based on this distance metric is one of the most commonly used methods.

Mathematically, the distance *D(X, Y)* between cluster X and Y is expressed as:

$$D(X,Y) = \min d(x,y) \tag{2.16}$$

$$x \in X, y \in Y,$$

Where,

$d(x,y)$ is the distance between objects x and y and

X and Y are two sets of objects (clusters).

- **Complete Linkage or Furthest-Neighbour Method (MAX):**

The dissimilarity between 2 groups is equal to the greatest distance between a member of cluster X and a member of cluster Y. This method tends to produce very tight clusters of similar cases. The distance between two clusters is computed as the maximal object-to-object distance between $\boldsymbol{d(x_i, y_j)}$ where objects $\mathbf{x_i}$ belong to the first cluster and objects $\mathbf{y_i}$ belong to the second cluster. This implies that, the distance between two clusters is computed as the distance between the two farthest objects in the two clusters. Complete link unlike the single link is less susceptible to noise and outliers and favours globular shapes.

Mathematically, it is expressed as:

$$D(X,Y) = \max d(x,y) \tag{2.17}$$

$$x \in X, y \in Y,$$

Where,

$d(x, y)$ is the distance between objects x and y;

X and Y are two sets of objects (clusters)

- **Average Linkage**

The distance or dissimilarity between two clusters is computed based on average distance between objects from the first cluster and objects from the second cluster. The averaging is performed over all pairs **(x, y)** of objects, where **x** is an object from the first cluster, and **y** is an object from the second cluster.

Mathematically, it is expressed as:

$$\frac{1}{Nx\ X\ Ny} \sum_{i=1}^{Nx} \sum_{j=1}^{Ny} d(xi, yi) \quad (2.18)$$

$$xi\ \in X, yi\ \in Y,$$

Where,

$d(x, y)$ is the distance between objects $x\ \in X$ and $y\ \in Y$;

X and Y are two sets of objects (clusters);

Nx and Ny are the numbers of objects in clusters X and Y respectively.

- **Group Average Linkage:**

The linkage function specifying the distance between two clusters is computed as the distance between the average values (the mean vectors or centroids) of the two clusters. In contrast to the single linkage and complete linkage approaches, where the distance is determined on the basis of extreme values of the distance function, this method considers the average between the distances computed between all pairs of patterns, one from each cluster. Hence, it is an intermediate technique between both approaches. Therefore, it requires more intensive computations.

Mathematically, it is expressed as:

$$D(X,Y) = \rho\left(\bar{x}, \bar{y}\right), \tag{2.19}$$

$$\bar{x} = \frac{1}{N_X}\sum_{i=1}^{N_X} \vec{x}\,i, \tag{2.20}$$

$$\bar{y} = \frac{1}{N_Y}\sum_{i=1}^{N_Y} \vec{y}\,i, \tag{2.21}$$

Where,

X and Y are two sets of objects (clusters);

N_X and N_Y are the numbers of objects in clusters X and Y respectively;

$\bar{x}$ and $\bar{y}$ are the mean vectors of the first and the second clusters, respectively;

$\rho\left(\bar{x}, \bar{y}\right)$ is the distance between vectors $\bar{x}$ and $\bar{y}$

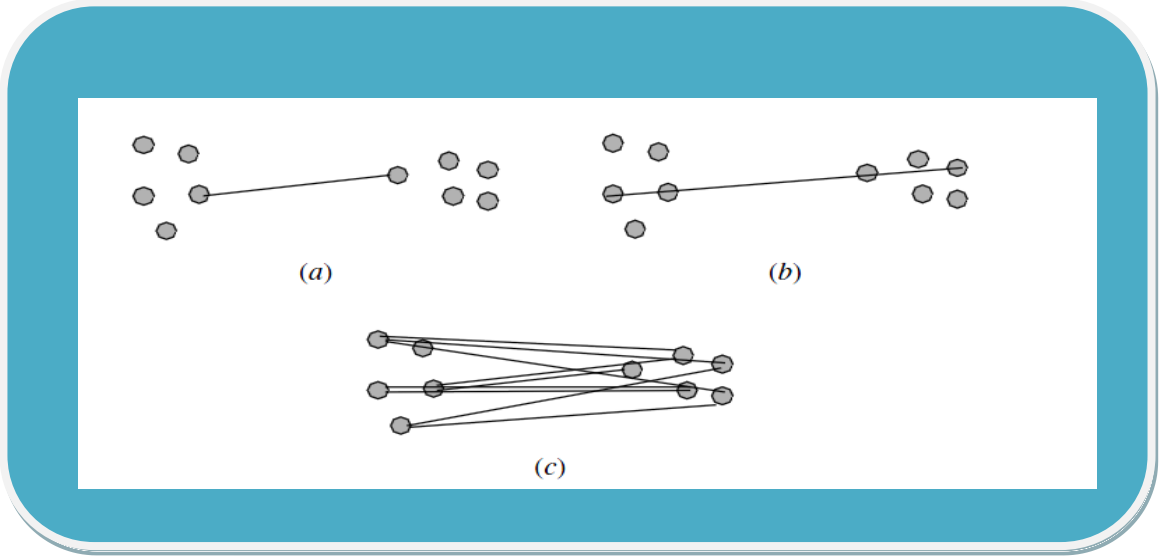

Figure 2-6: Two clusters and three main ways of computing the distance between them: (a) single link, (b) complete link, and (c) group average link.

- **Ward's Method:**

This method has much in common with analysis of variance (ANOVA). The linkage function specifying the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. Ward's Method seeks to choose the successive clustering steps so as to minimize the increase in ESS at each step. The ESS of a set X of N_x values is the sum of squares of the deviations from the mean value or the mean vector (centroid). For a set X the ESS is described by the following expression:

$$ESS(X) = \sum_{i=1}^{N_x} |xi - \frac{1}{Nx}\sum_{j=1}^{N_x} xj\,|^2 \qquad (2.22)$$

Where,

$|.|$ is the absolute value of a scalar value or the norm (the "length") of a vector.

Mathematically the linkage function - the distance between clusters X and Y is described by:

$$D(X,Y) = ESS(XY) - [ESS + ESS(Y)] \qquad (2.23)$$

Where,

XY is the combined cluster resulting from the fusion of clusters $X$and Y;
$ESS(.)$ is the error sum of squares described above.

Deciding on when to stop clustering will depend on two main criteria. Firstly, when clusters are too far apart to be reasonably merged, known as the distance criterion, and secondly, when there is a sufficiently small number of clusters referred to as the number criterion. The choice of decision criteria will depend on the purpose of the analysis.

2.6 Advantages and Disadvantages of Hierarchical Clustering

2.6.1 Advantages:

Hierarchical clustering algorithms are mostly used in practice (especially the agglomerative method) because it offers several advantages.

- Firstly, it has an embedded flexibility regarding the level of granularity. The dendogram shows a detailed structure of the data and clusters can be merged (agglomerative method) or split (divisive method) easily at any point of the tree. The smaller clusters that are generated may enable discovery.
- Secondly, this algorithm is capable of handling any form of distance or similarity measure and moreover, it is applicable to any type of attribute.

Despite these advantages, a couple of disadvantages still abound.

2.6.2 Disadvantages:

- The termination criteria are usually very vague and difficult to define in practice. For instance, it is not easy to determine the number of clusters which is ideal to obtain optimum results. This requires additional expert knowledge.
- Furthermore, it is not possible to undo intermediate clusters once they are constructed, hence improvement becomes difficult to implement. This means that, objects that are incorrectly grouped at an early stage cannot be undone.
- The use of different distance metrics for measuring distances between clusters may generate different results. It is therefore recommended to perform multiple experiments and compare the results in order to confirm the veracity of the original results.
- Another issue is the scaling problem. Most hierarchical clustering algorithms do not scale well. They usually have a time complexity of at least $O(n^2)$, where n is the total number of objects. The time complexity of an algorithm refers to the amount of time it takes to run, as a function of the size of the input to the problem.
- Finally, there is one main disadvantage in the interpretation of dendograms. Gordon (1994) stated that the absence of previous knowledge of the original data set often leads to misinterpretation by the human expert.

It should however be noted that linkage metrics presented in section 2.5 above based on the Euclidean distance usually have some restrictions. This is because they naturally predispose to clusters of proper convex shapes. However, this is not always the case with real data as they may exhibit various shapes and forms. In order to handle data of non-spherical shapes, wide variances in sizes and to resolve the problem of outliers, an alternative hierarchical clustering algorithm known as Clustering Using Representatives (CURE) was developed (Guha, Rastogi and Shim, 1998).
Experimental results obtained by the researchers confirmed that the quality of clusters produced by CURE is much better than those found by existing hierarchical algorithms. Furthermore, they demonstrated that random sampling and partitioning enable CURE not only to outperform existing algorithms but also to scale well for large databases without sacrificing clustering quality.

2.7 Partitional or Non-Hierarchical Clustering

In non-hierarchical clustering, data are divided into k partitions or groups with each group representing a cluster. In other words, it is the division of a set of data objects into non-overlapping subsets such that each data object is in exactly one subset. Therefore, unlike the hierarchical method, the number of clusters is to be known a priori. The main types under this category include the k-means, K-median and Fuzzy c-means.

2.7.1 K-means Clustering

This clustering technique is one of the most popular and simple approaches to clustering. Macqueen (1967) was the first person to use the term *k-means*. The goal is to partition *n* observations (objects) into *k* clusters such that each observation belongs to the cluster with the nearest mean. Generally, the algorithm is iterative and usually converges after several iterations to a local optimum. Unlike hierarchical clustering, the number of clusters *k* is stated apriori that is, the number of clusters is selected at the beginning, and the algorithm merely assigns each object to predetermined clusters.
The algorithm works by first selecting k locations at random to be the initial centroids for the clusters. Each observation is then assigned to the cluster which has the nearest centroid, and the centroids are recalculated using the mean value of assigned values. The algorithm then repeats this process until the cluster centroids do not change

anymore, or until they change less than a given threshold. As such, the K-means method is numerical, unsupervised, non-deterministic and highly iterative.
Below are the different steps in the k-means algorithm:

- Start with K randomly chosen points to define the centres of the K clusters,
- Assign each item to the closest point,
- Calculate the mean (centroid) of each cluster,
- Use the K means to define the centres of K new clusters and reassign each item to the cluster with the closest centre,
- Repeat the previous two steps until there is no change in the nature of the clusters between steps.

K-Means Algorithm has the following properties:

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

2.7.1a Advantages:

- Relatively simple clustering technique.
- K-means clustering may be computationally faster than hierarchical clustering when there are many variables and the number of clusters (k) is small.
- Furthermore, k-means tend to produce tighter clusters than hierarchical clustering especially if clusters are globular in shape; that is a cluster which is roughly spherical or convex in shape. This means that any line drawn between two cluster members or objects stays inside the boundaries of the cluster.

2.7.1b Disadvantages:

- K-means is not suitable for handling non-globular clusters or clusters of different sizes and densities. Non-globular clusters usually have very convoluted boundary and the mean is often irrelevant and may even lie outside the cluster.
- This clustering technique assigns objects to already predetermined k classes exclusively. This is not usually the case in reality as objects can belong to more than one cluster with varying degrees of membership probability.

- Moreover, by stating a fixed number of clusters a priori, it might become difficult to predict what k should actually be.
- Furthermore, it is not very efficient in clustering data that contains outliers, hence outlier detection and removal becomes necessary.
- Traditional k-means clustering was restricted to data for which there is a notion of centre (centroid). However, this has been resolved by using k-medoid which is more expensive to conduct.

Another variant of the K-means algorithm is the K-median. It differs from K-means in that it utilizes the manhattan or city block distance (equation 2.7) instead of the Euclidean distance (2.5) used in K-means. Therefore, it is less sensitive to outliers than the normal K-means due to the properties of the Manhattan distance metric (Bradley, Mangasarian and Street, 1997).

2.7.2 Fuzzy C-Means (FCM) Clustering

Standard clustering methods assume that data objects or points are exclusive and non-overlapping meaning that, each object is assigned to a single cluster only. However, there are some situations whereby an object can be rationally placed in more than one group or cluster implying they are overlapping or non-exclusive. For instance, a road crash can be caused by a factor related to human, vehicle or road infrastructure but reasonably, it can be attributed to a combination of these factors.

This is what fuzzy clustering is about. Fuzzy clustering reflects the fact that an object can simultaneously belong to more than one group or class depending on the criteria used. Based on this clustering method, every object belongs to every cluster with a known membership weight or probability that lies between 0 (absolutely does not belong) and 1 (absolutely belongs). The individual clusters are considered as fuzzy sets. Fuzzy sets are sets whose elements have degrees of membership. It was first introduced by Zadeh (1965) as an extension of the classical set theory. In the field of mathematics, a fuzzy set is one in which an object belongs to any set with a weight of between 0 and 1. However, in fuzzy clustering, an additional condition is imposed that the sum of the weights for each object must equal one.

Due to the fact that the membership weights for any object sum up to one, this technique does not address true multiclass situations. That notwithstanding, it is most

appropriate in avoiding the arbitrariness of assigning an object to only a single cluster whereas it can be quite similar to many. Therefore, it incorporates the existence of uncertainty in its analysis. However, in practical applications, fuzzy or probabilistic clustering is usually converted to an exclusive clustering by assigning each object to the cluster for which its membership weight is highest.

Fuzzy clustering is therefore similar to LC clustering. However, a major difference is that an object's grade of membership in fuzzy clustering are the 'parameters' to be estimated (Kaufman and Rousseuw, 1990) where as an individual's posterior class membership probabilities in LC clustering are computed from the estimated model parameters and its observed scores. This feature renders it possible to classify other objects belonging to the population from which the sample is taken which is not feasible with standard fuzzy clustering methods (Vermunt and Magidson, 2002).

This chapter gave an overview of widely used traditional clustering techniques. It is realised that despite the importance and popularity of these traditional clustering methods, they do have some major weaknesses especially as their computation rely heavily on a *distance* function. There also exist many other clustering methods including density and grid-based methods. Latent class models (LCM) and its application in clustering; Latent class cluster analysis (LCCA) which is the main analytical tool in this report is covered in the proceeding chapter.

Chapter 3: LATENT CLASS MODELS (LCM) AND LATENT CLASS CLUSTER ANALYSIS (LCCA)

The application of Latent class cluster analysis in the domain of traffic safety unlike standard clustering methods is still relatively new but a few studies do abound. Depaire Wets and Vanhoof (2008) showed that LCCA is an effective clustering technique for identifying homogenous traffic accident types. By applying the technique to a heterogeneous traffic accident data set, the researchers were able to segment the data into explicit clusters. In another study, Geurts, Wets, Brijs and Vanhoof (2003) used LCCA to cluster 19 central roads in the city of Hasselt (Belgium) into specific groups based on similar accident frequencies.

This chapter discusses latent class models and their application in clustering. The first part (section 3.1) begins with an explanation of latent class or finite mixture models. The statistical properties of the different functions are highlighted. This is then followed by a succinct step-by-step description of latent class cluster analysis in the next section.

3.1 Latent Class Models (LCM)

Latent class analysis was first introduced by Lazarsfeld (1950a, b) and Lazarsfeld and Henry (1968) as a technique for formulating latent attitudinal variables from dichotomous survey items. A latent variable, in contrast to an observable or manifest variable, refers to an unobservable construct which cannot be directly measured but can be inferred by a model from other variables that are observable. In this early model, it was assumed that the latent variable is categorical. Goodman (1974a, b) later on formalized and extended the methodology to nominal variables and is also credited with the development of the maximum likelihood (ML) algorithm that forms the basis of most LC software applications. LC models have been extended over the years and today include observable variables of mixed-scale type (nominal, ordinal, continuous and count) in the same analysis, the capacity to deal with sparse data, boundary solutions and other problem areas (Vermunt and Magidson, 2000). Moreover, for improved cluster or segment description the relationship between the latent classes and external variables can be assessed simultaneously with the identification of the

clusters by the inclusion of covariates. Therefore, the need for the usual second stage of analysis where a discriminant analysis is performed to relate the cluster results to demographic and other variables is eliminated. Apart from cluster analysis, LC modeling has also been applied to factor and regression analyses.

LC analysis can be viewed as a special case of model-based clustering for multivariate discrete data (Dean and Raftery, 2010). The main assumption of model-based clustering is that each observation comes from one of a number of classes, groups or subpopulations and each is modeled with its own probability distribution (McLachlan and Peel, 2000; Fraley and Raftery, 2002). Therefore, the overall population thus follows a finite mixture model (Fraley and Raftery, 2002) expressed as:

$$Y \sim \sum_{g=1}^{G} \pi g \; fg\,(Y) \tag{3.1}$$

Where,

$fg =$ the density function for group g

G = Number of groups

$\pi g =$ the mixture proportions,$0 < \pi g < 1$ and

$\sum_{g=1}^{G} \pi g = 1$

However, since in practice fg are from the same parametric family, as is the case with latent class analysis, the overall density is written as:

$$Y \sim \sum_{g=1}^{G} \pi g \; f\,(Y|\Theta g) \tag{3.2}$$

Where,

$\Theta g =$ set of parameters for the g^{th} group.

In latent class analysis, the variables are usually assumed to be independent given knowledge of the group an observation came from, an assumption referred to as local independence. A multinomial density is used to model each variable within each group

when the variables are categorical. The general density of a single variable x (with categories 1, . . . , d) given that it is in group g is then written as:

$$Y \mid g \sim \prod_{j=1}^{d} P_{jg}^{1\{y=j\}} \tag{3.3}$$

Where,

$1\{y = j\}$ is the indicator function equal to 1 if the observation of the variable takes value j and 0 otherwise, P_{jg} is the probability of the variable taking value j in group g, and d is the number of possible values or categories the variable can take.

Since we are assuming conditional independence, if there are k variables, their joint group density can be written as a product of their individual group densities. If $Y = y1, \ldots., yk)$, the joint group density is written as:

$$Y \mid g \sim \prod_{i=1}^{K} \prod_{j=1}^{d_i} P_{ijg}^{1\{y_i=j\}} \tag{3.4}$$

Where,

$1\{yi = j\}$ is the indicator function equal to 1 if the observation of the i^{th} variable takes value j and 0 otherwise, Pi_{jg} is the probability of variable i taking value j in group g and di is the number of possible values or categories the i^{th} variable can take.

More discussion of LC models is covered in Hagenaars and McCutcheon (2002), Clogg (1995) and McCutcheon (1987).

3.2 Latent Class Cluster Analysis (LCCA)

The application of latent class (LC) analysis as a clustering method has led to the development of Latent class cluster analysis (LCCA) in recent years. The first explicit connection between Latent class and cluster analysis was made by Wolfe (1970). Several terminologies have been used by researchers to describe such an application of LC analysis notably mixture likelihood approach to clustering (McLachlan and

Basford, 1988; Everitt, 1993), model-based clustering (Banfield and Raftery, 1993; Bensmail, Celeux, Raftery and Robert, 1997; Fraley and Raftery, 1998a, 1998b), mixture-model clustering (Jorgensen and Hunt 1996; McLachlan, Peel, Basford and Adams, 1999), Bayesian classification (Cheeseman and Stutz, 1995), unsupervised learning (McLachlan and Peel, 1996), finite mixture models (Fraley and Raftery, 2002) and latent class cluster analysis (Vermunt and Magidson, 2002).

3.2.1 LCCA for Continuous Variables

The basic LC cluster model is written as:

$$f(\mathrm{Yi}|\Theta) = \sum_{k=1}^{K} \pi_k \, f_k(Y_i|\Theta) \tag{3.5}$$

Yi denotes an object scores on a set of observed variables,

K is the number of clusters

π_k is the prior probability of belonging to latent class or cluster k or equivalently the size of cluster k.

As indicated in the above model, the distribution of Yi given the model parameters Θ, that is $f(\mathrm{Yi}|\Theta)$ is assumed to be a mixture of class specific densities, $f_k(\mathrm{Yi}|\Theta_k)$.

A great deal of research on LCCA has been done using continuous variables. These continuous variables are usually assumed to be normally distributed within the latent classes after applying an adequate non-linear transformation (Cheeseman and Stutz, 1995; Banfield and Raftery, 1993; McLachlan, 1988). Other distributions that can be used are the Student, Gompertz or Gamma distributions (McLachlan, Peel, Basford and Adams, 1999).

The general Gausssian distribution is the multivariate normal model with parameters μ_k and $\sum_k$. If no other restrictions are imposed, the LCCA problem involves estimating a separate set of means, variances and covariances for each latent class. The main goal in most applications is finding classes that differ with respect to their means or locations. The fact that the model allows classes to have different variances implies that classes may also differ with respect to the homogeneity of the responses to the

observed variables. The assumption is usually made in standard LC models with categorical variables that the observed variables are mutually independent within clusters. This assumption is not necessary in the case of continuous variables. The y variables (indicators or dependent or endogenous variables) may be correlated with clusters which may be cluster specific due to the fact that each class has its own set of variances. Therefore, the clusters differ with respect to their means and variances as well as to their correlations between the observed variables.

As the number of indicators and /or the number of latent classes increases, the number of parameters to be estimated also increases rapidly, especially the number of free parameters in the variance-covariance matrices $\sum_{k.}$ Hence, model restrictions that are imposed to obtain more parsimony and stability typically involve constraining the class-specific variance-covariance matrices.

A typical constraint model is the local independence model. It is obtained by making the assumption that all within-cluster covariances are equal to zero or in other words, the variance-covariance matrices, $\sum_k$ are diagonal matrices. Models that are less restrictive than the local independence model can be obtained by fixing some, but not all the covariances to zero that is, by assuming that certain pairs of y's are mutually independent within latent classes.

The equality or homogeneity of variance-covariance matrices across latent classes;

$\sum_k = \sum$ is another type of constraint. A homogenous or class-independent error structure of this nature yields clusters that have the same forms but different locations. These kinds of equality constraints can be applied in combination with any structure for $\sum$.

The reparameterising of the class specific variance-covariance matrices using the eigen value decomposition was proposed by Banfield and Raftery (1993) as follows:

$$\sum_k = \lambda_k D_k A_k D_K^T \tag{3.6}$$

Where;

$\lambda_k = |\sum_k|^{1/d}$ is a scalar; with d = number of observed variables

D_k = matrix with eigenvectors

A_k= diagonal matrix where elements are proportional to the eigen values of $\sum_k$ and is scaled such that $|A_k| = 1$.

An advantage of this model is that the three parameters each have a geometrical interpretation: λ_k indicates the volume of cluster k, D_k indicates its orientation and A_k is the shape. As such, restrictions that are imposed on these matrices can directly be interpreted in terms of the clusters. Generally, matrices are assumed to be class-independent and/or assigned simpler structures (diagonal or identity). An overview of many other specifications is covered in Fraley and Raftery (1998b).

Furthermore, the structure of the $\sum_k$ matrices can also be simplified using a covariance-structure model rather than a restricted eigen value decomposition. Many researchers have proposed the use of LC models in handling unobserved heterogeneity in covariance structure analysis (Arminger and Stein, 1997; Dolan and Van der Maas, 1997; Jedidi, Jagpal and DeSarbo; 1997). The same methodology can be used to restrict the error structure in LCCA with continuous indicators. Another important structure for $\sum_k$ which is related to the eigen value decomposition is the Factor analytic model (Yung, 1997; McLachlan and Peel, 1999). It is written as:

$$\sum_k = \Lambda_k \Phi_k + \cup_k \tag{3.7}$$

Where;

Λ_k = matrix with factor loadings

Φ_k = variance-covariance matrix of the factors

$\cup_k$ = diagonal matrix with unique variances

Restrictions can be imposed on this model by limiting the number of factors; for example to one and/or fixing some factor loadings to zero. These specifications render it possible to describe the correlations between the y variables within clusters or equivalently, the structure of local independencies by using a small number of parameters.

3.2.2 LCCA for Mixed-Mode Indicators

LC cluster models for continuous variables or indicators assume a restricted multivariate normal distribution for yi within each of the latent classes. In practice however, there exist many situations where there may be other variable types such as nominal, ordinal or count. LC models for nominal and ordinal variables assuming restricted multinomial distribution for the items are equivalent to standard exploratory LC models (Clogg 1981, 1995; Goodman 1974). LC models for Poisson count was postulated by Bockenholt (1993) and Wedel, DeSarbo, Bult and Ramaswamy (1993).

The specification of cluster models for indicators of different scale types is possible using the general structure of the LC model. This type of data is also referred to as mixed-mode data (Vermunt and Magidson, 2000; Lawrence and Krzanowski, 1996; Jorgensen and Hunt, 1996; Everitt 1988). Taking the local independence assumption into consideration, the LC cluster model is denoted as:

$$f(\mathrm{Yi}|\Theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{\mathrm{J}} f\mathrm{k}\,(\mathrm{y_{ij}}|\theta_{\mathrm{jk}}) \qquad (3.8)$$

Where;

J=total number of indicators

j = an individual indicator

The appropriate univariate distribution function for each element of $\mathrm{y_{ij}}$ of Yi is then specified, instead of specifying the joint distribution of Yi given class membership using a single multivariate distribution. Relevant distributions for continuous $\mathrm{y_{ij}}$ include the univariate normal, student, gamma and log-normal distributions. Multinomial distribution (restricted) is the normal choice for discrete nominal or ordinal variables. Finally, Poisson, binomial and negative binomial are used for count variables.

It is assumed in the above equation (3.8) that the y's are conditionally independent within the latent classes. However, it is possible to relax this assumption by using the relevant multivariate rather than univariate distributions for sets of locally dependent y variables. Presenting a separate formula for this situation is unnecessary as the index

in equation 3.8 can be considered to denote a set of indicators and not just a single indicator. The multivariate normal distribution is used for a set of continuous variables and a set of nominal /ordinal variables can merge into a (restricted) joint multinomial distribution. A multivariate Poisson model can be used to model correlated counts. The specification of the mixed multivariate distributions is rather complex. Two possible ways of modeling the relationship between a nominal /ordinal and a continuous y was proposed by Krzanowski (1983):

- Conditional Gaussian where the categorical variable can be used as covariate in the normal model and
- Conditional multinomial distribution where the continuous variable is used as a covariate in the multinomial model.

The conditional Gaussian distribution in LC clustering with combination of categorical and continuous variables was used by Hunt and Jorgensen (1999) and Lawrence and Krzanowski (1996). Moreover, local dependencies with a Poisson variable are handled in the same way; by allowing its mean to be dependent on the relevant continuous or categorical variables.
The inclusion of local dependencies between indicators is very important in LCCA for two main reasons:

- It ensures that the final solution does not have so many clusters. By including a few direct effects between y variables, a simple solution with fewer clusters is obtained. However, allowing within-cluster association is also disadvantageous in that relevant clusters may remain hidden due to direct effects.
- Moreover, a better classification of objects into clusters is possible when the local independence assumption is relaxed. Two variables that are assumed to be locally independent imply that they likely contain some overlapping information that should not be used when determining the class membership of an object. Furthermore, if a significant bivariate dependency is omitted from a LC cluster model, the subsequent locally dependent indicators tend to get a very high weight in the classification formula (equation 3.11) compared to the other indicators.

3.2.3 Inclusion of Covariates

The general LCC modeling approach deals with mixed-mode data and allows for many different specification of the (correlated) error structure. An important extension of this model is the prediction of class membership by the inclusion of covariates. It is logical to differentiate (endogenous) variables that serve as indicators of the latent variable from (exogenous) variables that are used to predict to which cluster an object belongs. This is the same idea expressed by Clogg (1981) in LCM with external variables.
In some applications, the latent cluster variable can be used as a predictor of an observed response variable rather than as a dependent variable. By using the response variable as one of the y variables, a model in which cluster variables serve as a predictor can be obtained.
Using the same standard structure as in equation 3.8 above, the LC cluster model becomes:

$$f(\mathrm{Yi}|\mathrm{Zi},\Theta) = \sum_{k=1}^{K} \pi_{k|\mathrm{Zi}} \prod_{j=1}^{J} f_k\,(y_{ij}|\theta_{jk}) \tag{3.9}$$

where;
Zi =covariate values of object i

Synonyms for z's are exogenous variables, external variables, concomitant variables, grouping variables or inputs. To reduce the model complexity (number of parameters), of the probability of belonging to class k given covariate values Zi, $\pi_{k|\mathrm{Zi}}$ is generally restricted by a multinomial logit model; that is a logit model with "*linear effects*" and no higher order interactions.
Conceptually, a more general specification is obtained when covariates are assumed to have direct effects on the indicators:

$$f(\mathrm{Yi}|\mathrm{Zi},\Theta) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}|\mathrm{Zi}} \prod_{\mathrm{j}=1}^{\mathrm{J}} f_{\mathrm{k}}\left(\mathrm{y_{ij}}|\mathrm{Zi},\theta_{\mathrm{jk}}\right) \tag{3.10}$$

In this case, the conditional mean of y variables can now be directly related to the covariates. As such, the implicit assumption in the previous specification (equation 3.9) that the influence of z's on y's goes completely through the latent variable is relaxed.

There is also another possibility to have direct effects of z's on y's by using a simple short-cut to specify direct effects between indicators of different scale types. This is obtained by using one of the two variables involved both as a covariate (no influence on class membership) and as an indicator.

3.2.4 Model Estimation

The maximum likelihood (ML) and the maximum posterior (MAP) estimators are the main methods use in estimating the parameters of the various types of LC cluster models. The log-likelihood function required in ML and MAP approaches are derived from the probability density function defining the model. Bayesian MAP estimation involves maximizing the log-posterior distribution which is the sum of the log-likelihood function and the logs of the priors for the parameter. Both methods are quite similar, but ML estimate has an edge over MAP in that it prevents occurrence of boundary or terminal solutions; that is probabilities and variances cannot become zero. Given a very small amount of prior information, the parameter estimates are forced to stay within the interior of the parameter space. Typical examples of priors are the Dirichlet priors for multinomial probabilities and the inverted-Wishart priors for the variance-covariance matrices in multivariate normal models.

The expectation-maximization (EM) algorithm (McLachlan and Krishnan, 1997) or some modification of it is used in most statistical software packages to compute the ML and MAP estimates. Generally, the algorithm starts with a number of EM iterations and switches to Newton-Raphson (NR) when it is close enough to the final solution. In this manner, the advantages of both algorithms are combined as EM is very stable even when far from the optimum and NR has higher speed when close to the optimum.

The occurrence of a local solution is a common problem in LC analysis. An ideal way to avoid ending with a local solution is to use multiple sets of starting values. Most

computer programs for LC clustering such as Latent Gold have automated the search for good starting values using several sets of random starting values as well as solutions obtained with other cluster methods.

In LC cluster analysis, the interest does not only lie in the estimation of the model parameters. The classification of objects into clusters is also a significant "*estimation*" problem and is based on posterior class membership probabilities as follows:

$$\pi_{k|Yi,Zi} = \frac{\pi_{k|Zi} \prod_j f_k(yi|Zi,\Theta_{jk)}}{\sum_k \pi_{k|Zi} \prod_j f_k\,(yi|Zi,\Theta_{jk)}} \tag{3.11}$$

Modal allocation is the standard classification method; that is assigning each object to the class with the highest posterior probability.

3.2.5 Model Selection

Model selection is a major research topic in LC clustering. The two main issues deal with decisions about the number of clusters and the form of the model given the number of clusters.

Standard likelihood ratio tests between nested models are used to make assumptions with respect to the forms of the clusters given their number. A typical example is between a model with an unrestricted covariance matrix and a model with restricted covariance matrix. Wald tests can be used to assess the significance of certain included variables while the Lagrange multiplier tests are used to assess that for excluded terms. However, these kinds of chi-squared tests cannot be used to determine the number of clusters (Vermunt and Magidson, 2002).

There are three major model selection tools in LC clustering using the information criteria notably the Bayesian information criterion (BIC), Akaike information criterion (AIC), and the Consistent Akaike information criterion (Fraley and Raftery, 1998a).

In the LCA computer program, Latent Gold (Vermunt and Magidson, 2000), these criteria are reported in the output and computed in two ways:

- Based on the likelihood-ratio chi-squared statistic (L^2) and degrees of freedom (df) as follows:

$$\mathrm{BIC}_{L2} = L^2 - \log(\mathrm{N})\mathrm{df}, \tag{3.12}$$

$$\mathrm{AIC}_{L2} = L^2 - 2\,\mathrm{df}, \tag{3.13}$$

$$\mathrm{CAIC}_{\mathrm{L2}} = L^2 - [\log(\mathrm{N}) + 1]\,\mathrm{df}. \tag{3.14}$$

With
L^2 = likelihood-ratio chi-squared statistic,
df = degrees of freedom.

- Based on the log-likelihood ($\log \mathcal{L}$) and the number of estimated parameters (npar) defined as:

$$\mathrm{BIC}_{\log \mathcal{L}} = -2 \log \mathcal{L} + (\log N)\,\mathrm{npar}, \tag{3.15}$$

$$\mathrm{AIC}_{\log \mathcal{L}} = -2 \log \mathcal{L} + 2\,\mathrm{npar}, \tag{3.16}$$

$$\mathrm{CAIC}_{\log \mathcal{L}} = 2 \log \mathcal{L} + [(\log N) + 1]\,\mathrm{npar}. \tag{3.17}$$

With
$\log \mathcal{L}$ = log-likelihood
npar = number of estimated parameters

These statistical values are a measure of how well the model describes the data (model fit) and also the complexity of the model in terms of number of parameters. Lower values of BIC, AIC and CAIC indicate a better model in terms of parsimony.

Other computationally intensive methods have been developed recently including parametric bootstrapping (McLachlan, Peel, Basford and Adams, 1999) and Markov Chain Monte Carlo methods (Bensmail Celeux, Raftery and Robert, 1997) to determine the number of clusters and their forms. A fully automated model selection method using approximate Bayes factors, different from BIC was proposed by Cheeseman and Stutz (1995).

There are other techniques for evaluating LC cluster models which are based on the uncertainty of classification or in other words, the separation of the clusters. Apart from the estimated total number of misclassifications, other measures can be used to indicate how well the indicators predict class membership. Some of these indices combine information on model fit and information on classification errors. Examples of such indices include the classification likelihood (C) and the approximate weight evidence (AWE) in Celeux, Biernacki and Govaert (1997).

Chapter 4: METHODOLOGY

This chapter describes the methodological aspects of this study. It discusses the research strategy and the empirical techniques that were applied. The data sources, data manipulation and transformations are discussed in the first section. The second part explains the data analytical technique of Latent Class Cluster Analysis (LCCA) and how it is modeled in Latent Gold.

4.1 Data Description and Preparation

The data used in this project is obtained from the Directorate-general Statistics Belgium. It is a record of road crashes that occurred in Belgium in the year 2005.The registration of accidents involving casualties in Belgium is carried out by police officers at the accident site using a form which is designed for this purpose known as the "*Analysis Form for Traffic Accidents with Casualties*". This implies that only injury accidents are recorded as those involving only material damage are not taken into consideration. The final database contains other important attributes related to the accident including the vehicle type, road infrastructure, environmental conditions and above all the characteristics of the road users involved.

The original data set is available in Microsoft Excel format on two separate sheets labeled *"Accidents"* and *"Victims"*. The *Accident* spreadsheet has 40,563 observations (accidents or crashes) and that for *Victims* has 88,645 crash victims. The *Accident* sheet was then exported into SAS 9.2, where the necessary transformations were performed as elaborated in the proceeding paragraphs.

As the purpose of this thesis is to analyse road crashes at intersections, the first step involves obtaining a sample data set on accidents that occurred at intersections. This variable is coded as a binary response with "1" indicating accidents on intersections and "2" indicating accidents that occurred on other segments of the road rather than intersections. By setting the original variable coded as r4 to 1, that is where r4 = 1, the sample data set of accidents that occurred at intersections is obtained. This analysis takes only the first collision into consideration to avoid complications that can arise from combining single collision crashes with multiple collision crashes. For example, a single collision crash can be erroneously treated as a missing value for the second collision type whereas in reality there was no second collision. Furthermore,

incorporating variables like age and gender could be done by extracting these from the *"Victims"* sheet. However, merging both sheets (Accidents and Victims) produced a data set containing details about the victims. This is because each accident identification (id) in the *"Accidents"* sheet is linked to several victims (rows) in the *"Victims"* sheet. Since the focus is to cluster individual accidents or crashes and to avoid complications that can arise from complex adjustments, the decision was made to leave out these variables. That notwithstanding, the other variables including the collision type, road type, time, season, weather conditions, pedestrian or moped involvement amongst others provide a vivid description of the crashes at intersections.

The result indicates that out of 40 563 cases of accidents, 13 953 took place at intersections. This implies that about 34% of registered road crashes in Belgium in the year 2005 occurred at intersections. This high proportion vividly portrays the strategic role of intersections on road safety and hence the importance of a research of this nature. In order to facilitate interpretation, the labels of the different columns were changed to reflect their actual description. The variables of interest were then selected to be used for subsequent analysis. A list of the selected variables is displayed on table 4-6 on page 64.

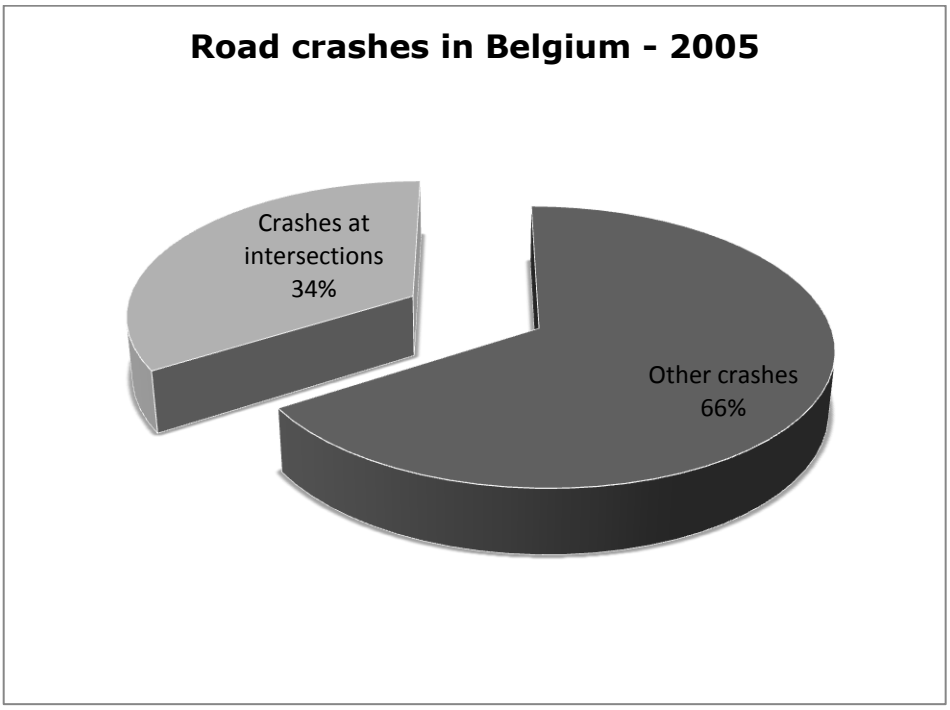

Figure 4-1: Road crashes at intersections in Belgium - 2005

Furthermore, it was necessary to alter the categories of certain variables in order to incorporate other possible scenarios. The following variables were transformed to facilitate the analysis.

- Traffic control in the centre of the intersection (TRAFFIC_C)

This variable indicates the type of traffic control at the centre of the intersection. The original data set contained 5 categories for this variable as displayed on table 4-1a directly below. This is our main variable that will be used to segment the data to distinguish among the intersection types.

Table 4-1a: Traffic control in the centre of the intersection (TRAFFIC_C)

1	traffic policeman
2	functioning three-coloured traffic lights
3	defective three-coloured traffic lights or amber flashing light
4	Right-of-way signs B1 or B5
5	right of way to traffic from the right

However, there are two possibilities that can occur at the intersection. If the first position has a value of 3 (defective three-coloured traffic lights or amber flashing light), either option 4 (right-of-way signs B1 or B5) or 5 (right of way to traffic from light) will be present. For the others, if the first road segment contains values 1, 2, 4 or 5, then the second position is left blank. In order to incorporate these, the categories were adjusted as follows:

- Categories 3 and 4 are combined to give a new category labeled 6 (An instance where there was a defective three-coloured traffic lights or amber flashing light and the right-of-way sign B1 or B5 was present).
- Categories 3 and 5 are also combined to form a new category labeled 7 (An instance where there was a defective three-coloured traffic lights or amber flashing light and there was right of way to traffic from the right).

Two new categories, 6 and 7 were therefore created and category 3 in the data set was eliminated. The final variable had the following categories:

Table 4-1b: Adjusted traffic control at intersection

1= traffic policeman
2= functioning three-coloured traffic lights
4= right-of-way signs B1 and B5
5= right of way to traffic from the right
6= defective three-coloured traffic lights or amber flashing light and the right-of- way sign B1 or B5 was present
7= defective three-coloured traffic lights or amber flashing light and there was right of way of traffic from the right

- The variable *HOUR*, which indicates time of day (0h-23h), was divided into six broad categories as shown on table 4-2 below, in order to capture variations in crashes over different periods of the day. Furthermore, the grouping was necessary to ensure that a reasonable number of observations are recorded for the specified periods without delving into too much detail which might inhibit interpretation when hourly periods are considered.

Table 4-2: Time of the day (*HOUR*)

Hours	**Classification**
06 – 09	Morning
10 – 12	Late morning
13 – 15	Afternoon
16 – 18	Evening rush
19 – 21	Evening
22 – 05	Night

- Months were aggregated and a new variable *SEASON* was created. In the original data set, months were denoted by values of 1-12 representing January to December. The aggregation was done as was the case with *HOUR* in order to ensure that a reasonable number of observations are available for each period since the data pertains to just a single year. Moreover, observing seasonal variations in accidents despite the inherent detailed information loss can still help to gain some useful

information. The months were aggregated as depicted below on table 4-3. Therefore, in the final analysis, the variable *Month* was replaced by *SEASON.*

Table 4-3: Aggregation of months into seasons (*SEASON*)

Months	**Seasons**
Dec.[12], Jan.[01] and Feb.[02]	Winter
March[03], April[04] and May[05]	Spring
June[06], July [07] and August[08]	Summer
Sept.[09], Oct.[10] and Nov.[11]	Autumn

- The variable *road condition* (Road_C) was also adjusted to reduce the number of categories. Initially, it had six categories and two categories; dry (1) and clean (4) roads were merged to form one category. Hence, five categories are used in further analysis.

Table 4-4: Adjusted categories for road condition (*Road_C*)

Values	**Classification**
1	Dry and Clean
2	Wet puddles
3	Black ice, snow
4	Dirty (sand, gravel, leaves etc.)
5	Unknown

The number of categories was also reduced for the variable *Weather*. Initially eight categories were present and three categories namely rainfall (2), snowfall (5) and hailstorm (6) were all merged under one broad category; precipitation. Fog (3) was also merged with other for example dense smoke (7) to bring the total number of categories to four.

Table 4-5: Adjusted categories for Weather conditions (*Weather)*

Values	Classification
1	Normal
2	Precipitation (rainfall, snowfall and hailstorm) including strong wind or gusts.
3	Fog or other e.g dense smoke
4	Unknown

Below are the other selected variables. The complete list of selected variables is shown on table 4-6, pages 64.

4.1.1 Other selected variables

- **Collision type** (***COL_T***)
  This depicts the various types of collisions that occurred at the intersection. There are six categories for this variable. Multiple collisions and single driver collisions are not included. As mentioned earlier, only the first collision is taken into consideration.
- **Road types (first and second): *Road_T1* and *Road_T2***
  This variable denotes the road type outside of the intersection where the accident occurred. It has three categories; 1 = motorway, 2 = regional road and 3 = municipal road.
- **Speed limit difference (*SL_D*)**
  This indicates the difference in speed limit between the two roads at the intersection. It is coded as a binary variable with 0 = no difference and 1 = difference in speed limit.
- **Light conditions (*LightC*)**
  This indicates the lighting conditions available when the accident occurred. It has five categories ranging from daylight to darkness as indicated on table 4-6 below.
- **Built-up area (*BUA*)**
  This indicates whether the accident occurred in an agglomeration or not. A value of *"1"* indicates the accident was inside a built-up area and *"2"* implies it was outside of a built-up area.

- **Ignored red light or does not give priority to right of way (Behav1)**
  This indicates whether the driver violated the red traffic light or priority rule or not before the crash. It is also a binary variable with 0 = No and 1 = Yes.
- **Pedestrian Involvement**
  This indicates whether a pedestrian was involved in the crash or not. It is a binary response with 0 = No and 1 = Yes.
- **Involvement of mopeds and cyclists (MOP_C)**
  This also indicates whether a moped rider or bicyclist was involved in the crash; 0 = No and 1 = Yes.
- **Number of fatalities (*FATS*)**
  This variable denotes the number of fatalities (or deaths) that resulted from the crash.
- **Seriously injured crash victims (*SER_INJ*)**
  This variable shows the number of victims who sustained serious injuries.
- **Slightly injured crash victims (*SLI_INJ*)**
  This variable shows the number of crash victims who sustained slight or minor injuries.

Table 4-6 List of selected variables

Variable	Values
Collision type *(COL_T)*	1= between drivers, head-on collision or when crossing each other, 2 = between drivers, rear-end collision or side collision, 3 = between drivers, side collision, 4 = between a driver and pedestrian, 5 = collision with an obstacle (on and off the road), 6= Other or unknown
Traffic control at intersection *(TRAFFIC_C)*	1= traffic policeman, 2= functioning three-coloured traffic lights, 4= right of way signs B1 and B5, 5= right of way to traffic from the right, 6= defective three-coloured traffic lights or amber flashing light and the right of way signs B1 or B5 was present, 7= defective three-coloured traffic lights or amber flashing light and there was right of way of traffic from the right
Road_T1* and *Road_T2	The road type outside the intersection. 1= motor way, 2= regional road, 3= municipal road or other.
SL_D	Speed limit difference. Whether there was a speed limit difference between the roads. 0 = No, 1 = Yes.
Season *(SEASON)*	1= Winter, 2= Spring , 3= Summer, 4= Autumn
Built-up area *(BU_A)*	The location of accident. 0 = inside built-up area, 1 = outside built-up area.
Light conditions *(LightC)*	1= Day, 2= Dawn, twilight, 3= Night with public lighting, 4= Night with no public lighting, 9= Unknown
Road_C	Road conditions. 1= dry and clean, 2 = wet, puddles, 3 = Black ice, snow, 4 = dirty (e.g sand, gravel, leaves e.t.c), 5 = unknown.
Driver behaviour (Behav1)	Ignored red light or does not yield to right of way. 0 = No, 1 = Yes.
Time of day (HOUR)	1= Morning, 2 = Late morning, 3 = Afternoon, 4 = Evening rush, 5 = Evening, 5 = Night
Weekend	Whether an accident occurred during the week end. 0 = No, 1 = Yes
Weather	1 = Normal, 2 = Precipitation (rainfall, snowfall and hailstorm) and strong winds or gusts, 3 = Fog and other e.g dense smoke, 4 = Unknown
Pedestrian	Whether a pedestrian was involved in the crash or not. 0 = No, 1 = Yes.
MOP_C	Whether a moped rider or bicyclist was involved in the crash. 0 = No, 1 = Yes.
FATS	The number of fatalities resulting from the accident.
SER_INJ	The number of crash victims with serious injuries.
SLI_INJ	The number of crash victims with slight or minor injuries.

4.1.2 Data Segmentation

Let us label the sample data set, containing the selected variables of all crashes that occurred on intersections as *CRASH*. This is the final sample data set prior to the partition into different smaller samples based on the variable *Traffic control in the centre of the intersection* denoted by *TRAFFIC_C*.

The main goal of this research is to segment or partition crashes that occurred at intersections into groups of distinct crash types. To achieve this objective, the variable TRAFFIC_C (traffic control in the centre of the intersection) is used to repartition the final data set obtained labeled *CRASH*. It should be recalled that this variable consists of six different categories (or intersection types). This procedure was performed in SAS 9.2 and the following connotations on table 4-7 are used to depict these newly created sample data sets. The sas codes used for performing this and other data manipulations are provided in Appendix 1.

Table 4-7: Sample data sets based on intersection type

Data set	Intersection type
Intersec_TP	Intersection with a Traffic policeman.
Intersec_TL	Intersection with functioning three-coloured traffic lights.
Intersec_RS (B1 or B5)	Intersection with right-of-way signs B1 or B5.
Intersec_TR	Intersection with right of way to traffic from the right.
Intersec_DTLRS	Defective three-coloured traffic light or amber flashing light and the right-of-way sign B1 or B5 was present.
Intersec_DTLTR	Defective three-coloured traffic light or amber flashing light and there was right of way to traffic from the right.

4.2 Research Method

Before the data implementation in latent Gold, basic exploratory data analytical techniques are used to gain some prior knowledge about the data. Frequency tables and charts are constructed for the different variables. This is then followed by the application of LCCA in Latent Gold for the different sample data sets.

4.2.1 Latent Class Cluster Analysis (LCCA)

Traditional clustering methods like the K-means and hierarchical clustering make use of unsupervised classification algorithms that group objects or items that are close together according to an ad hoc definition of distance (Vermunt and Magidson, 2000). The prime objective of clustering is to discover hidden structures or patterns within data by creating homogenous clusters or groups. However, these distance-based approaches have several shortcomings as elaborated in the proceeding section. Due to the inherent weaknesses of these early approaches, it is advisable to utilize advanced statistical clustering techniques like LCCA.
LCCA is a model based clustering technique that overcomes the weaknesses of distance-based clustering methods. It has several advantages over the earlier clustering methods as elucidated below.

4.2.2 Advantages of LCCA over traditional Clustering Methods

Latent class cluster analysis has several properties that make it superior to traditional clustering methods. The LC analysis is a model-based approach that utilizes estimated membership probabilities or weights estimated directly from the model to classify objects into appropriate clusters. Hence, the selection of a distance measure is unnecessary. This is more advantageous over the traditional clustering techniques that group similar objects based on a ad hoc definition of *'distance'*. However, different distributions are needed depending on the variable type. The normal Gaussian distribution is used when the descriptive feature is continuous, a Multinomial distribution is applied for a nominal variable, an adjacent-category Logistic regression model is selected for an ordinal variable and the Poisson distribution is appropriate for count variables (Vermunt and Magidson, 2005).

Furthermore, there are well defined statistical criteria to select the number of clusters and other model features. Information criteria including BIC, AIC and CAIC are powerful model selection tools. Moreover, LC models do not rely on traditional modeling assumptions such as linear relationship and homogeneity which are usually violated in practice. As such, they are less prone to biases when data do not conform to these assumptions.

Furthermore, variables of mixed scale type (nominal, ordinal, continuous or count) or any combination of these can be incorporated into the same model. Added to this, in

order to ensure an improved cluster description of the latent (unobserved) and indicator (manifest) variables, covariates can be included in the analysis which is not possible with traditional clustering algorithms.
It is possible using LC analysis to build models from large data sets since it does not require large memory demands (Brijs, 2002). This is further facilitated by the development of high-speed computers and sophisticated algorithms.

This model-based clustering technique is also advantageous over standard distance-based cluster analysis in that no decisions have to be made regarding the scaling of the observed variables; for example, when dealing with normal distributions with unknown variances, the results will be unaltered whether the variables are normalized or not. There is usually a scaling problem when traditional non-hierarchical methods are used. In LC clustering, both simple and complex distributional forms can be used for the observed variables within the clusters thereby making it highly flexible.

To conclude, LCCA is more realistic as objects are assigned membership weights of varying probabilities of belonging to different clusters unlike standard clustering with the exception of fuzzy clustering which assigns objects to a single cluster.

4.2.3 Implementation of LCCA in Latent Gold and Data processing

There are many computer software packages used for estimating various types of LC models. These include; NORMIX (Wolfe, 1970), Autoclass (Cheeseman and Stutz, 1995), Classmix (Moustaki, 1996), LEM (Vermunt, 1997), MCLUST (Fraley and Raftery, 1998b), Mcplus (Muthen and Muthen, 1998), EMMIX (McLachlan, Peel, Basford and Adams, 1999), MULTIMIX (Hunt and Jorgensen, 1999) and above all Latent Gold (Vermunt and Magidson, 2000). These computer programs differ in several aspects including the types of cluster models they implement (multivariate normal distribution/ and or mixed-mode data), possibility to include covariates in the model, estimation method used, Algorithm (Expectation Maximisation-EM or Newton Raphson-NR) utilized and the type of source code and the operating environment.
Latent Gold just like LEM is a full Windows based program which facilitates user interaction. It is capable of handling all variable types and the specification of the structure of the error-covariance matrices is less complex. Furthermore, its multiple sets of starting values greatly reduce the possibility of occurrence of a local solution and it is easy to detect local dependencies to be included in the model using bivariate residual

measures. Due to these extra capabilities, Latent Gold was used to conduct the main analysis in this research.

The latent class model for mixed mode data; equation 3.8 earlier stated in chapter 3, section 3.2.2 (page 49) is implemented in Latent Gold taking into consideration the assumption of local independence within clusters. Due to this assumption, the parametric complexity of the model is reduced and facilitates the combining of descriptive variables of various scale types.

The individual data sets were then exported from SAS 9.2 and saved in SPSS format (.sav extension) which is readable in Latent Gold. One of the data sets, *Intersec_DTLRS* had no observations (traffic crash or accident) and two others *Intersec_TP* and *Intersec_DTLTR* had fewer observations to be effectively clustered, so only three were left for further analysis in Latent Gold. Seventeen variables were used to build the models as displayed on the "*Variable tab*" in Latent Gold in figure 4-2 below. The variables on the right, with their scale types specified, indicate those that were actually included in the models. These data sets are stored in the folder named Traffic_Con in two formats: SPSS format (.sav extension) and Excel format (.csv extension).

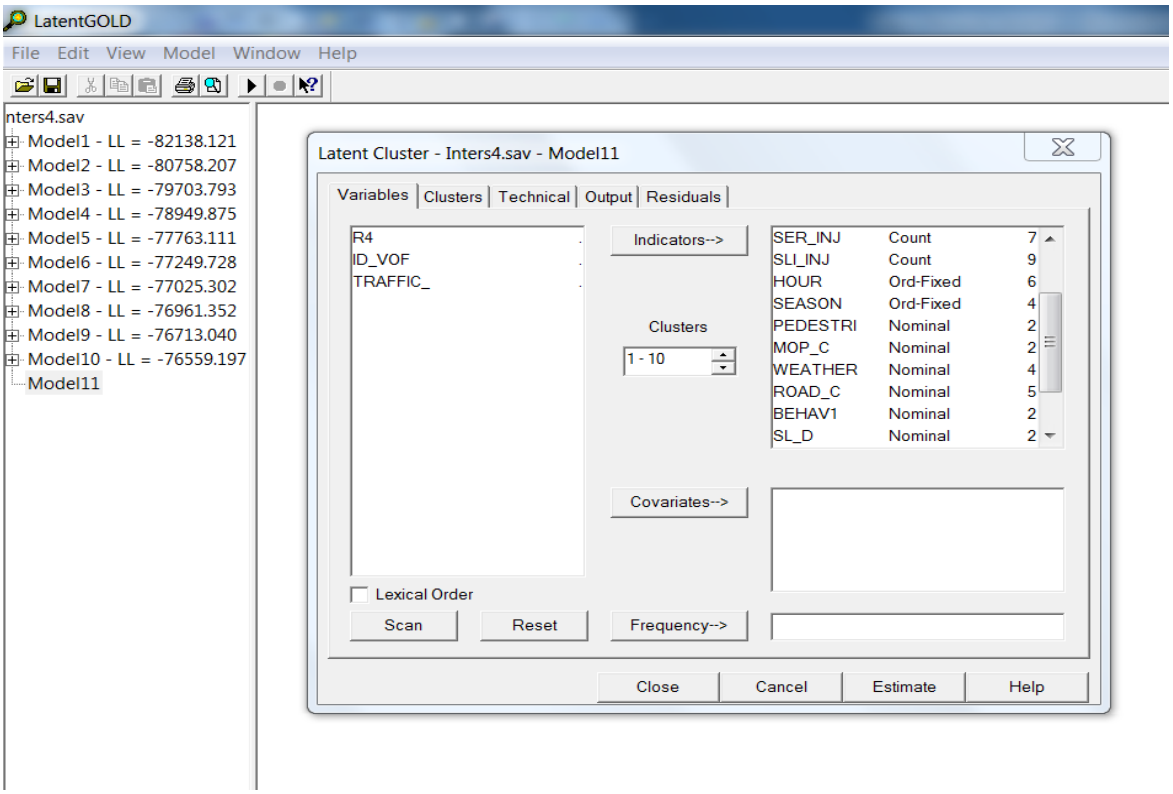

Figure 4-2: Screen shot of the variable tab in Latent Gold.

Models are built in Latent Gold by running several iterations. The variables used to estimate the models (see table 4-6 on pp.62) were selected and their scale types specified (figure 4-2).The program by default automatically selects random starting values and models can be built from a single cluster to several clusters by specifying the desired number of clusters. For each data set, a range was selected from 1 to 10 and the best models were chosen based on their BIC, AIC and CAIC values. Cluster analysis was conducted for three intersection types namely intersection with functioning three-coloured traffic lights (*Intersec_TL*), intersection with right-of-way sign B1 or B5 (*Intersec_RS*) and intersection with right-of-way to traffic from the right (*Intersec_TR*).

Chapter 5: ANALYSIS OF RESULTS AND DISCUSSION

The results derived from descriptive statistics are presented in the first section (section 5.1). This gives us an insight into the data structure before the main statistical analysis is performed. It is then followed by results of LCCA in Latent Gold for each of the three data sets (section 5.2). A description of the clusters or crash types is further elaborated. Finally, the crash types identified at the various intersection types are compared for the different data sets.

5.1 Results of Basic Exploratory Data Analysis (EDA)

Frequency tables and charts (graphs) are used to describe the variables as they were mostly nominal and ordinal variables. The results are presented in the proceeding subsections.

5.1.1 Crashes at intersection types

The final data set labeled *CRASH*, which contains all the road crashes that occurred at intersections, is partitioned based on the manner of traffic control at the intersection (or intersection types). The six data sets obtained are analyzed using frequency tables and charts. The results are displayed on table 5-1 and figure 5-1 below.

Table 5-1: Frequency distribution of crashes at various intersection types.

Data set	Frequency	Percent (%)	Cumulative Frequency	Cumulative %
Intersec_RS	6968	51.58	6968	51.58
Intersec_TR	3529	26.12	10497	77.7%
Intersec_TL	2747	20.33	13244	98.04
Intersec_DTLTR	204	1.51	13448	99.55
Intersec_TP	61	0.45	13509	100.00
Intersec_DTLRS	0	-	13509	100.00

Frequency Missing = 444

It is clear from the above frequency distribution table (table 5-1) and the chart below (figure 5-1) that more than half the overall number of crashes at intersections (51.58 %) occurred at intersections with a right-of-way sign B1 or B5 present (*Intersec_RS)*. *Intersec_TR* follows next with 26.12% while *Intersec_TL* is comprised of 20.33% of crashes at intersections. The lowest number of crashes (except *Intersec_DTLRS* with zero observation) was recorded at intersections where a traffic policeman was present (*Intersec_TP*) with just 0.45% of the overall crashes while the remainder (1.51%) occurred at Intersec_*DTLTR*. It should be noted that no accident was recorded at intersections where there was a defective three-coloured traffic light or amber flashing light and a right-of way-sign B1 or B5 was present (*Intersec_DTLRS*). The observations (the number of crashes or accidents) at the various intersection types make up the size of the individual data sets. In a nutshell, almost all crashes (98.04%) occurred at intersections types *Intersec_RS, Intersec_TR* and *Intersec_TL*. There were 444 missing values in the data set. This figure indicates the number of accidents at intersections where the specific intersection type was not mentioned.

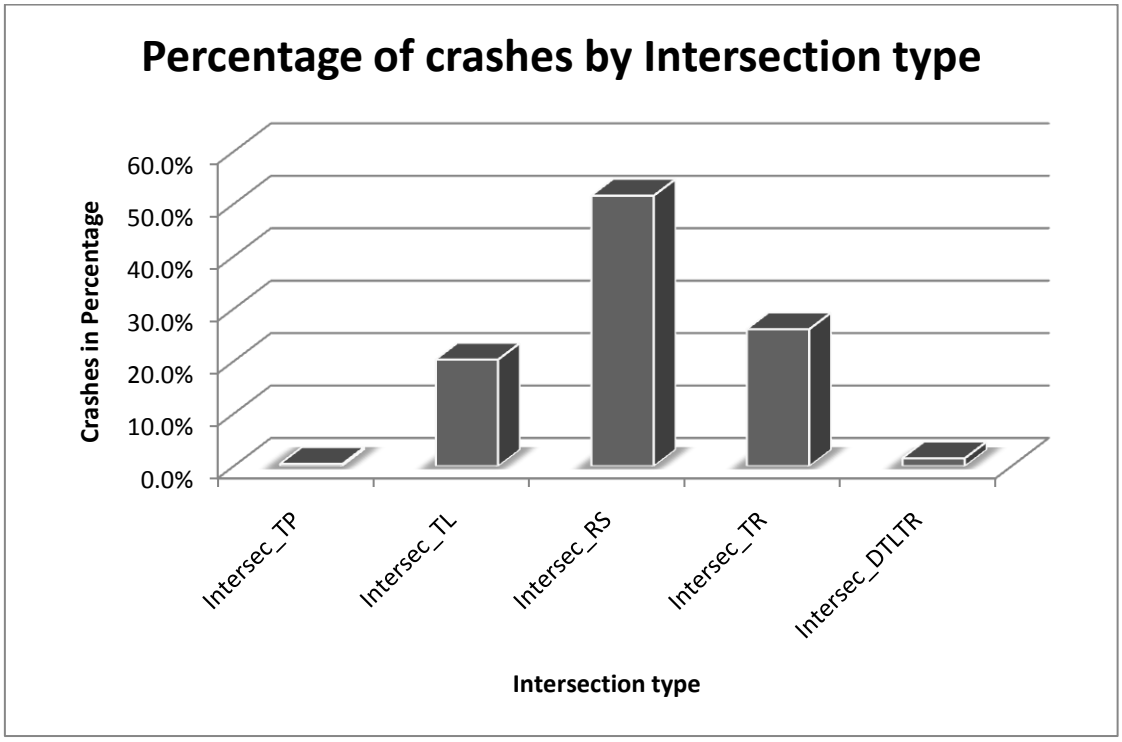

Figure 5-1: Proportion of crashes at various intersection types.

In order to gain an insight into the severity of crashes at the different types of intersections, the severity of crashes notably the number of fatalities (*FATS*), seriously injured victims (*SER_INJ*) and slightly injured victims (*SLI_INJ*) was computed and compared among the different data sets. This information is presented in Table 5-2 and figure 5-2 below.

Table 5-2: Severity of crashes based on intersection types

	Interse_TP		**Intersec_TL**		**Intersec_RS**		**Intersec_TR**		**Intersec_DTLTR**	
Outcome	Values	%	Values	%	Values	%	Values	%	Values	%
FATS	1	1.32	36	0.89	114	1.18	22	0.48	2	0.62
SER_INJ	7	9.20	319	7.89	973	10.11	351	7.68	12	3.69
SLI_INJ	68	89.47	3687	91.22	8534	88.70	4198	91.84	311	95.70
Number of victims	76		4042		9621		4571		325	

Legend

FATS-Fatalities

SER_INJ - Seriously injured victims

SLI_INJ - Slightly injured victims

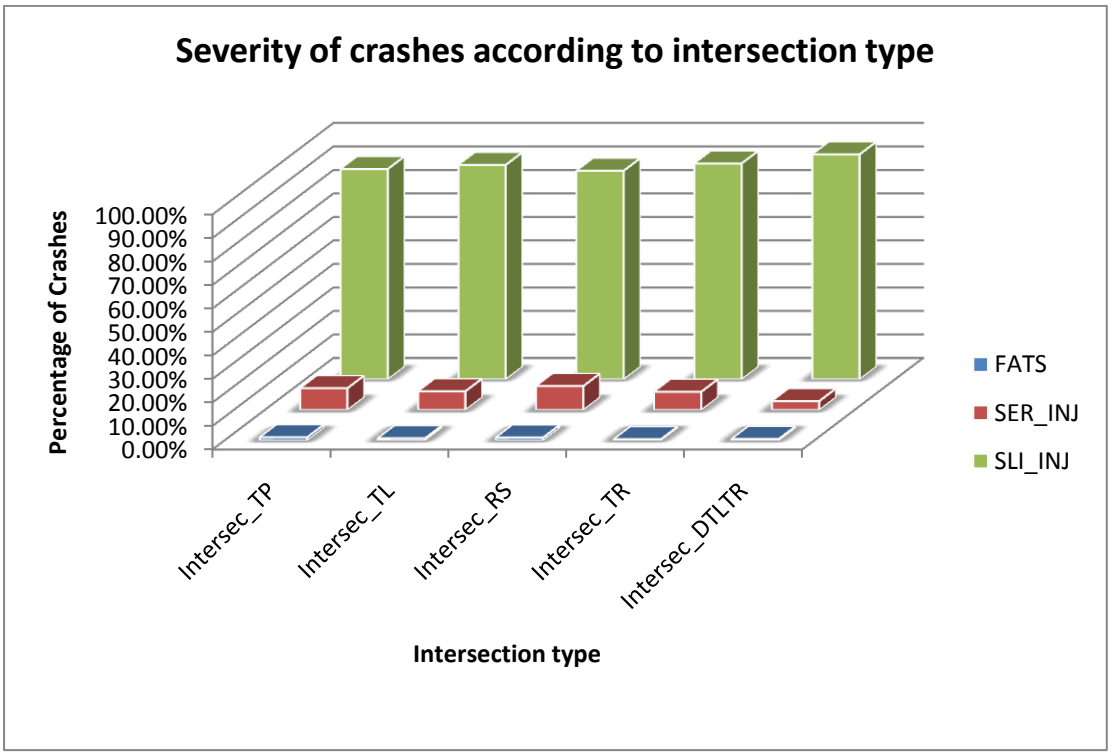

Figure 5-2: Severity of crashes based on intersection type.

A total of 18 635 victims were involved in crashes at intersections with 175 fatalities, 1 662 serious cases and the remainder being slightly injured. In other words, 0.94% of crash victims sustained fatal injuries, 8.92% sustained serious injuries and 90.14% were slightly injured.

Intersection with a traffic policeman (*Intersec_TP*) surprisingly has the highest percentage of crash victims, who were killed, with a fatality rate of 1.32% followed by intersection with a right-of-way sign B1 or B5 (*Intersec_RS*) with 1.18%. However, *Intersec_TP* has the lowest number of crash victims. The lowest percentage of deaths was registered at *Intersec_TR* with 0.48%. In terms of serious injuries, *Intersec_RS* is the highest (10.11%) followed by *Intersec_TP* (9.2%) and *Intersec_TL* (7.89%), *Intersec_TR* (7.68) while *Intersec_DTLTR* has the lowest percentage with 3.69% and also the highest percentage of victims who sustained slight injuries.

5.1.2 Collision types (COL_T)

The frequency distribution was also computed for the collision type. The results are displayed below on tables 5-3 and 5-4 and figure 5-3. More than half the number of collisions (57%) involved side collision between drivers (*COL_T3*). Rear-end collision between drivers (*COL_T2*) made up 15% followed by head-on or collision while crossing between drivers (14.40%). Collision with a pedestrian (*COL_T4*) and that with an obstacle (*COL_T5*) had the same proportion of accidents of close to 6%.

Table 5-3: Frequency distribution of Collision type (COL_T)

COL_T	Frequency	Percent (%)	Cumu. Freq.	Cumu. %
3	7 836	57.14	7 836	57.14
2	2 120	15.46	9 956	72.60
1	1 975	14.40	11 931	87.01
4	781	5.70	12 712	92.7
5	781	5.70	13 493	98.4
6	220	1.60	13 713	100

Missing = 240

The collision types were later on compared with the severity levels or outcome of the crashes to give a clear notion of their level of riskiness. The results are shown on table 5-4 below. From the analysis, it was revealed that the highest fatality rate was recorded in collisions with an obstacle both on and off the road combined (*COL_T5*) with a fatality ratio of 2.7%. The next highest is collision between a driver and pedestrian with a rate of 2% (*COL_T4*) followed by side collision with 0.9% fatality rate. Head-on collision (*COL_T1*) was the penultimate with 0.7% and the least was rear-end collisions with 0.4%.

In terms of serious injuries, collision with an obstacle (*COL_T5*) and collision with a pedestrian (*COL_T4*) were the highest making up 16% of the crash victims originating from the respective collision types. Head-on collision was the third with 10% of seriously injured victims followed by side collisions (*COL_T3*) with 8.5%. Rear-end collisions again had the lowest share of seriously injured crash victims.

Consequently, rear-end collisions had the highest proportion of slightly injured casualties with 95% followed by side collisions with 91%. Hence, driver-pedestrian collisions (*COL_T4)* and collision with an obstacle (*COL_T5*) both have the lowest share of slightly injured crash victims with 82% each. Finally, 1.6% of crashes representing 220 crashes were classified as other or unknown.

Table 5-4: Collision type and severity of crashes (%).

Outcome	**COL_T1**		**COL_T2**		**COL_T3**		**COL_T4**		**COL_T5**		**COL_T6**	
	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%
FATS	19	0.7	12	0.4	98	0.9	17	2	27	2.7	2	0.8
SER_INJ	282	10	156	5	912	8.5	134	16	157	16	14	5.5
SLI_INJ	2612	89	2923	94.6	9757	91	696	82	808	82	239	94
Total	2913		3091		10767		847		992		255	

Collision types

COL_T1 = between drivers, head-on collision or when crossing each other.

COL_T2 = between drivers, rear-end collision.

COL_T3 = between drivers, side-collision.

COL_T4 = between a driver and pedestrian

COL_T5 = Collision with an obstacle (off and on the road).

COL_T6 = Other or unknown

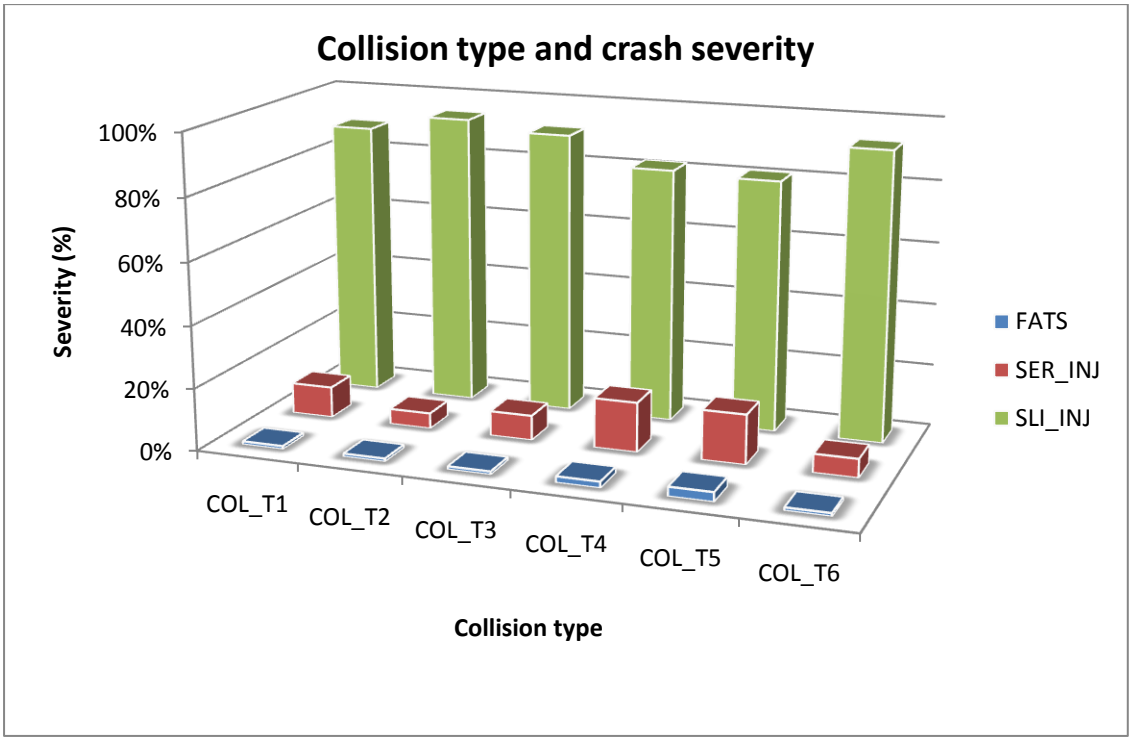

Figure 5-3: Collision type and crash severity

5.1.3 Crashes according to time of the day (*HOUR*)

Table 5-5 and figure 5-4 below show the frequency distribution for the variable *HOUR* which indicates the proportion of accidents during different periods of the day. Most accidents were recorded during the evening rush hour (25.67%), followed by during the afternoon and morning periods with 18.98% and 17.38% respectively and 16% during the late morning period. Close to 80% of accidents were registered during these four periods. There was one missing value.

Table 5-5: Number of crash victims according to time of the day (*HOUR*)

HOUR	Frequency	Percent (%)	Cumulative Frequency	Cumulative %
Evening rush (16–18)	3 581	25.67	3 581	25.67
Afternoon (13–15)	2 648	18.98	6 229	44.65
Morning (10–12)	2 425	17.38	8 654	62.03
Late morning (6–9)	2 234	16.01	10 888	78.04
Evening (19–21)	1 652	11.84	12 540	89.88
Night (22–5)	1 412	10.12	13 952	100.00

Frequency Missing = 1

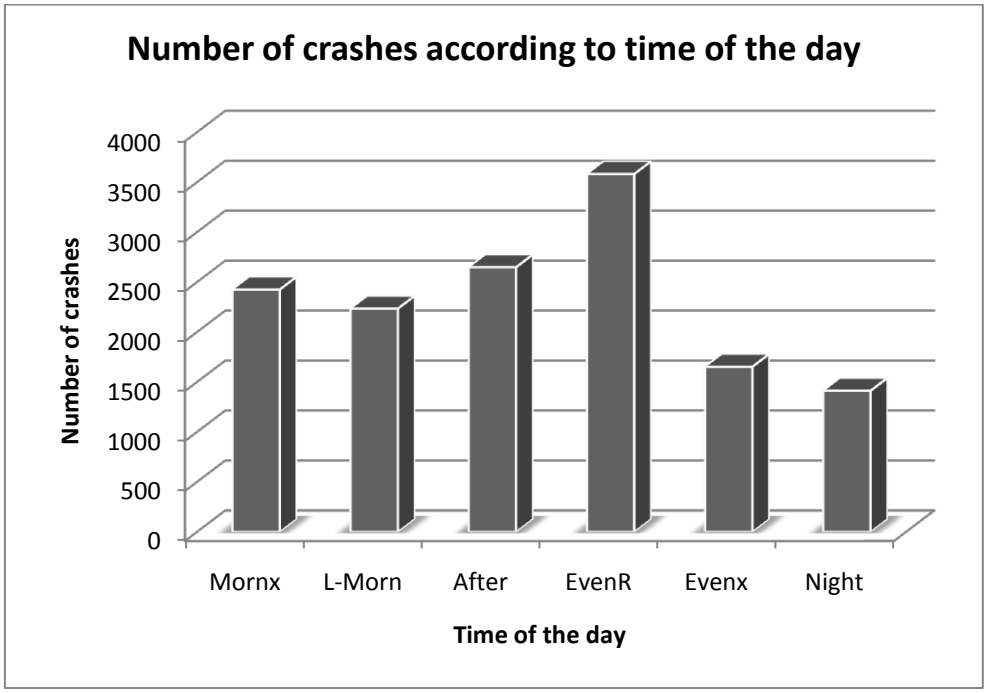

Figure 5-4: Number of crashes according to time of the day (*HOUR*).

5.1.4 Crashes during the Week-end

About 24% of crashes occurred during the week end (Saturday and Sunday) while the rest occurred during week days.

Table 5-6: Number of crashes during the week (end)

Weekend	Frequency	Percent (%)	Cumulative Frequency	Cumulative %
No (0)	10 5272	75.77	10 572	75.77
Yes (1)	3 381	24.23	13 953	100

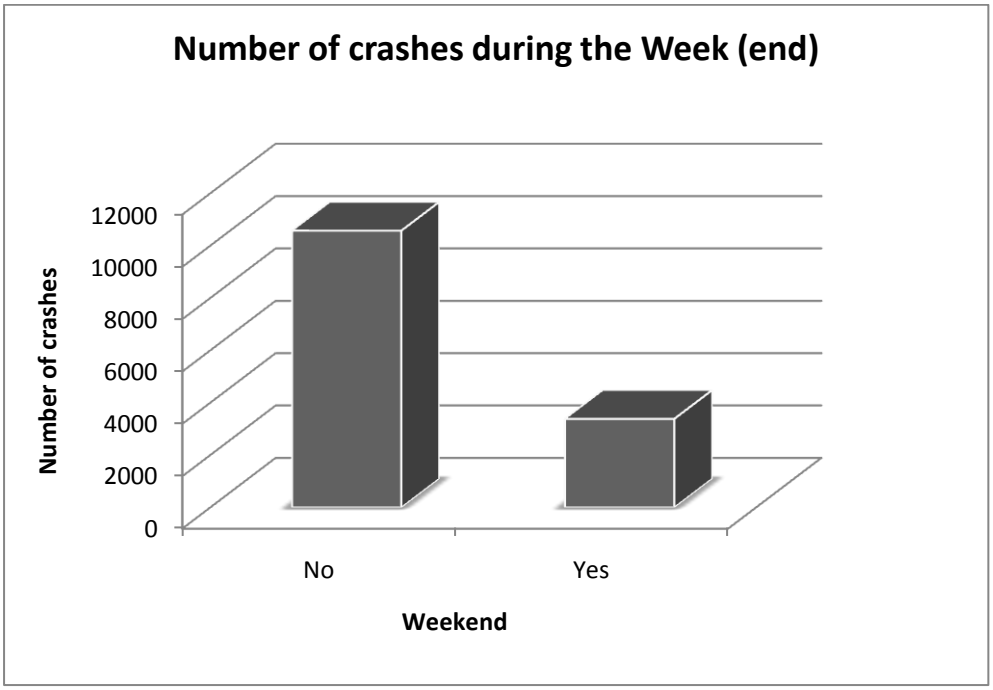

Figure 5-5: Number of crashes during the week (end)

5.1.5 Seasonal distribution of crashes (*SEASON*)

It is also important to describe the seasonal variations in crashes. As indicated on table 5-7 and the chart below (figure 5-6), the highest number of crashes occurred during autumn with 27.92% of the crashes. Summer and spring had roughly the same proportion with 25.86% and 24.70% respectively. The winter season was the lowest and made up 21.53%. Generally, the number of accidents did not fluctuate considerably throughout the different seasons.

Table 5-7: Seasonal distribution of crashes (*SEASON*)

SEASON	Frequency	Percent (%)	Cumulative Frequency	Cumulative %
Autumn	3 895	27.92	3 895	27.92
Summer	3 608	25.86	7 503	53.77
Spring	3 446	24.70	10 949	78.47
Winter	3 004	21.53	13 953	100.00

Figure 5-6: Number of crashes according to season (*SEASON*).

5.1.6: Crashes according to built-up area (*BU_A*)

To have an insight about the concentration of accidents, a frequency distribution was calculated for *BUA.* As depicted on table 5-8 and figure 5-7, 63% of accident victims were recorded inside built-up areas while the rest occurred outside.

Table 5-8: Number of crashes according to built-up area (*BUA*)

BU_A	Frequency	Percent (%)	Cumulative Frequency	Cumulative Percent
Inside	8 779	62.93	8 779	62.93
Outside	5 172	37.07	13951	100.00

Frequency Missing = 2

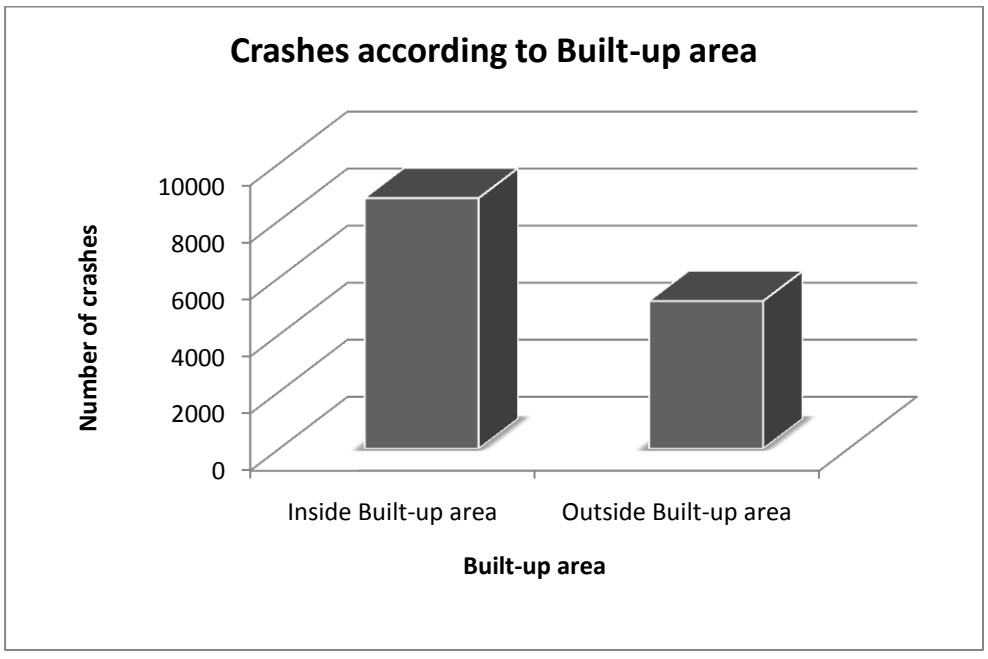

Figure 5-7: Number of crash victims according to built-up area (*BUA*)

5.2 Additional description of Intersec_TP and Intersec_DTLTR

An extra descriptive analysis was conducted for *Intersec_TP* and *Intersec_DTLTR* because they had few observations and could not be effectively modeled in Latent Gold.

5.2.1 Intersection with a Traffic policeman (Intersec_TP)

Collisions at Intersec_TP were dominated by side-collisions (46.67%) followed by rear-end collisions making up 20%. Collision with an obstacle (*COL_T5*) was the least with 5%. This trend is similar to the overall crashes at intersections (See Appendix 3.1a and 3.1b).

Regarding the time of the day, most crashes occurred during the afternoon period (26%) and Evening rush hour (25%) while the least was at night making up 7%. (See Appendix 3.2a and 3.2b)

Considering the days of the week, 30% of crashes occurred during the week-end (Saturday and Sunday) while the rest occurred on week days. This percentage is slightly higher than the overall case. (See appendix 3.3a and 3.3b).

Close to 50% of crashes occurred during the autumn season, 28% during summer and the lowest in winter with 10%. (See appendix 3.4a and 3.4b).

Of the sixty-one crashes that occurred at intersections with a traffic policeman (*Intersec_TP*), eight involved pedestrians making up 13%. (See Appendix 3.5a and 3.5b)

5.2.2: Intersection with a defective traffic light and right of way to traffic from the right (Intersec_DTLTR)

Side collisions made up 64% of crashes at this intersection type followed next by rear-end collisions making up 19% and the least was collision with an obstacle which was 3%. (See appendix 4.1a and 4.1b).

The majority of crashes occurred during the evening rush hour with 24% followed by the morning period with 20% and the least was at night with 11%. (See appendix 4.2a and 4.2b).

Crashes during the week-end made up 29%. This is slightly higher than the overall rate. (See appendix 4.3a and 4.3b).

Most crashes also occurred during the autumn season with 30% followed by summer with 27%. The least number of crashes occurred in spring making up 19% (See appendix 4.4a and 4.4b).

Out of the 204 crashes that occurred at this intersection type, 10 involved pedestrians making up (See appendix 4.5a and 4.5b).

5.3 Results of LCCA in Latent Gold

The results from analyzing the individual data sets in Latent Gold are presented in this section. The analysis is presented per data sets and the best models are selected based on the three afore-mentioned selection criteria notably the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC). These statistical criteria asses the model fit and also take into account the model complexity in terms of number of parameters. Models with lower values are better than those with higher values when comparing models.

In addition, the entropy criterion (denoted as I) is also used to measure how well the model predicts class membership thereby assessing the quality of the estimated clustering solution. This value ranges between 0 and 1, with values closer to 1 indicating a better classification of observations into their specific clusters or groups.

This is computed as:

$$I(k) = 1 - \frac{\sum_{i=1}^{n}\sum_{z=1}^{k} pik \ln(pik)}{n \ln (1/\mathrm{k})} \qquad 5.1$$

Where,

pik is the posterior probability that case i belongs to cluster k

and $pik \ln(pik) = 0 \; if \; pik = 0.$

However, this classification criterion is already included in the output in Latent Gold.

As mentioned before, Intersection with a traffic policeman (*Intersec_TP*) and intersection with a defective three-coloured traffic light or amber flashing light and there was a right of way to traffic from right (*Intersec_DTLTR*) could not be effectively clustered in Latent Gold as they had fewer number of crashes. They both had 61 and 204 crashes respectively. Nonetheless, the three other intersection types are the most important as they registered about 98% of the overall crashes at intersections. Below are the clustering results obtained for these intersection types.

5.3.1 Intersection with functioning three-coloured Traffic lights (*Intersec_TL*)

This data set comprises crashes that occurred at intersections with a functioning three-coloured traffic. It contains a total of 2747 traffic crashes making up about 20% of crashes at intersections and it is the third largest sample data set. The 6-cluster model was chosen based on low BIC, AIC and CAIC values with an entropy criterion, $I(5)= 0.84$ implying a good classification . Figure 5-8 below displays the evolution of the three model selection criteria when adding clusters. The profile output from Latent Gold is provided in Appendix 2.1.

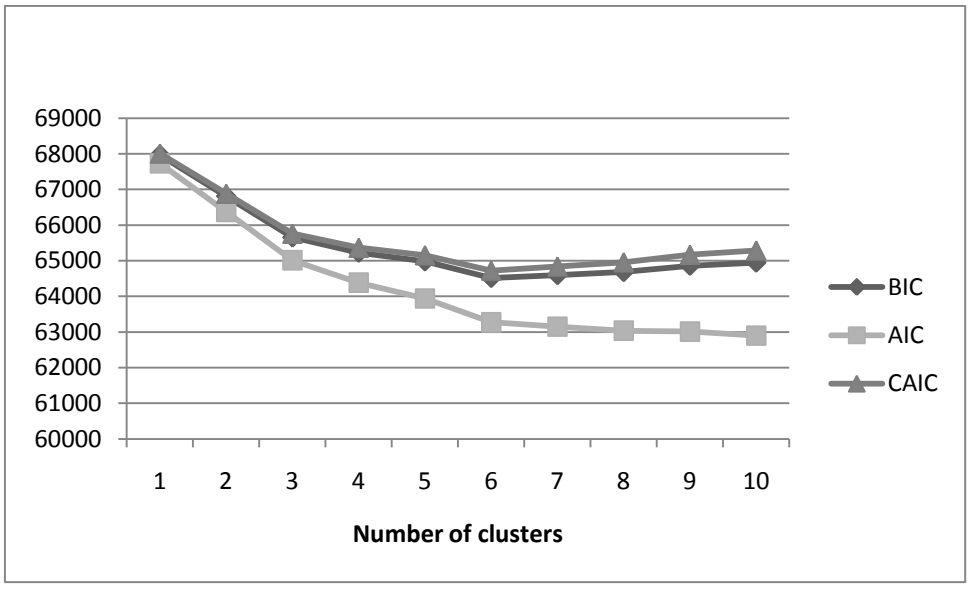

Figure 5-8: Evolution of BIC, AIC and CAIC with the addition of clusters (Intersec_TL)

The first cluster mostly contains side collisions (60%) that occurred during the evening rush hour (26%) in autumn (31%) usually in normal weather conditions (97%) and road

conditions (93%) inside built-up areas (83%). This cluster is referred to as *"road crashes in normal weather and road conditions during the day".*

Cluster two mainly include side-collisions (72%) during the evening rush hour (25%) where there was a significant speed limit difference between the roads outside the intersections (53%) usually outside of built-up areas (88%) and 73% of the drivers ignored the red light. This cluster overlaps with the first cluster except that there was no difference in speed limit between the roads and unlike cluster two; the crashes were concentrated inside built-up areas and they were fewer traffic light violations. These features clearly differentiate it from the other clusters. This cluster is called *"road crashes with traffic light violation outside built-up areas".*

Cluster three is comprised of rear-end collisions (82%) that occurred during the winter season (26%) in normal weather (98%) with virtually no red light violation (99.9%) and no pedestrian involved (100%). It is distinct from the other clusters as it is the only cluster that is dominated by rear-end collisions. This cluster is called *"road crashes with rear-end collision with no pedestrian casualty".*

The fourth cluster involves side collisions (41%) that occurred almost entirely at night (99%) with public lighting (95%) during the winter season (28%). It is the only cluster that is dominated by night crashes. This cluster is referred to as *"road crashes at night with public lighting".*

The fifth cluster contains crashes that occurred during precipitation (69%) and consequently on wet roads with puddles (89%) during the winter season (30%). It overlaps with cluster four except that unlike cluster four, most of the crashes occurred during the day. This cluster will be labeled "*road crashes during precipitation in winter*".

The last cluster comprises collisions with pedestrians (98%) reflected by the high involvement of pedestrian casualty (99.9%) inside of built up areas (91%). Therefore it is called "road *crashes with collisions between drivers and pedestrians*".

Table 5-9: Crash types at Intersec_TL

Cluster	Crash type	Size (%)
1	Road crashes in normal weather and road conditions during the day.	33
2	Road crashes with traffic light violation outside built-up areas.	18
3	Road crashes with rear-end collision with no pedestrian	16

	casualty.	
4	Road crashes at night with public lighting.	13
5	Road crashes during precipitation in winter.	13
6	Road crashes with collisions between drivers and pedestrians	7

5.3.2 Intersection with a right-of-way sign B1 or B5 (*Intersec_RS*).

This data set contains crashes that occurred at intersections with a right-of-way sign B1 or B5.This is the largest data set with a total of 6968 traffic crashes making up more than 50% of the overall number of crashes at intersections. The 7-cluster model was chosen based on the aforementioned selection criteria. The entropy criterion $I(6)$ equals 0.86 which indicates a good classification of crashes into clusters. Changes in BIC, AIC and CAIC values as the number of clusters was increased are shown on the graph in figure 5-10 below. The profile output from Latent Gold is provided in Appendix 2.2.

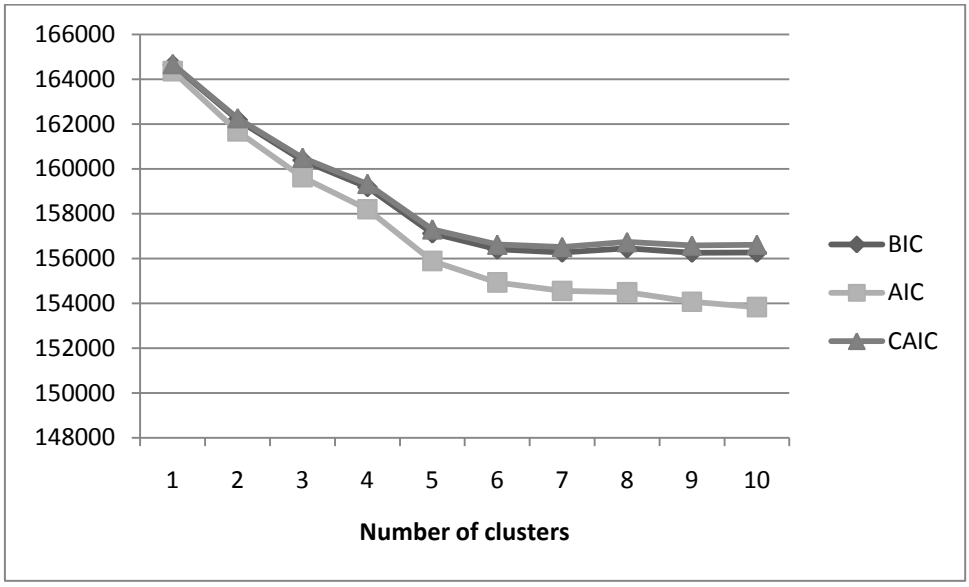

Figure 5-9: Evolution of BIC, AIC and CAIC with addition of clusters (Intersec_RS)

Cluster one consists of side collisions (79%) during the evening rush hour (30%) in autumn (31%) usually in normal weather (97%) and road conditions (93%) inside of

built-up areas (100%) and 51% of drivers did not respect the yield sign (B1) or the stop sign (B5). It is referred to as "*road crashes in normal weather and road conditions".*

Cluster two also involves mainly side-collisions (83%) during the evening rush hour (30%) in autumn (29%) with no pedestrian involvement and at least one road user failed to yield to right-of-way or stop sign (64%) mostly outside of built-up areas (93%). It overlaps with cluster one except that the crashes mainly took place outside of built-up areas. This cluster is referred to as *"road crashes with side-collisions when a driver failed to respect a B1 (yield) or B5 (stop) sign".*

The third cluster mainly comprises crashes involving rear-end collisions (62%) during the evening rush hour (31%) and there was almost no violation of right of way signs (99.9%). It is quite similar to cluster two except the collision type and violation of right of way. It is called *"road crashes involving rear-end collisions outside of built-up areas".*

Cluster four depicts road crashes involving side-collisions (62%) during precipitation (63%) and consequently on wet roads with puddles (89%) during summer (29%). These two features clearly distinguish it from the other clusters. It is referred to as *"road crashes during precipitation in summer".*

The fifth cluster comprises road crashes that occurred at night (98%) with public lighting (93%) during the autumn (27%) and summer (27%) seasons. This differs from the other clusters as it contains mainly night accidents. It is called *"road crashes at night with public lighting".*

The sixth cluster contains collisions between drivers and pedestrians (97%) hence; it is composed almost entirely of pedestrians (99.9%) inside built-up areas (87%). This cluster is labeled as *"road crashes with collisions between drivers and pedestrians".*

The last cluster is quite unique from the others in that the roads outside the intersections were principally regional roads unlike the others which were mostly municipal roads. This cluster is comprised of road crashes during evening rush hour (29%) in the summer season (28%) outside of built-up areas (91%) on intersecting regional roads. This cluster is referred to as "road crashes on intersecting regional roads outside built-up areas".

Table 5-10: Crash types at Intersec_RS

Cluster	Crash type	Size (%)
1	Road crashes in normal weather and road conditions.	30
2	Road crashes with side-collisions when a driver failed to respect a B1 (yield) or B5 (stop) sign.	27
3	Road crashes involving rear-end collisions outside of built-up areas.	15
4	Road crashes during precipitation in summer.	13
5	Road crashes at night with public lighting.	9
6	Road crashes with collisions between drivers and pedestrians.	4
7	Road crashes on intersecting regional roads outside built-up areas.	2

5.3.3 Intersection with right of way to traffic from the right (*Intersec_TR*)

This data set contains crashes that occurred at intersections with a right of way to traffic from the right. It is the second largest data set with 3529 observations. The model with five clusters was chosen as the ideal since it has the lowest BIC value. The corresponding entropy criterion is 0.88 indicating a good classification. Figure 5-11 below displays fluctuations in the BIC, AIC and CAIC values as the number of clusters were added. The profile output from Latent Gold is provided in Appendix 2.3.

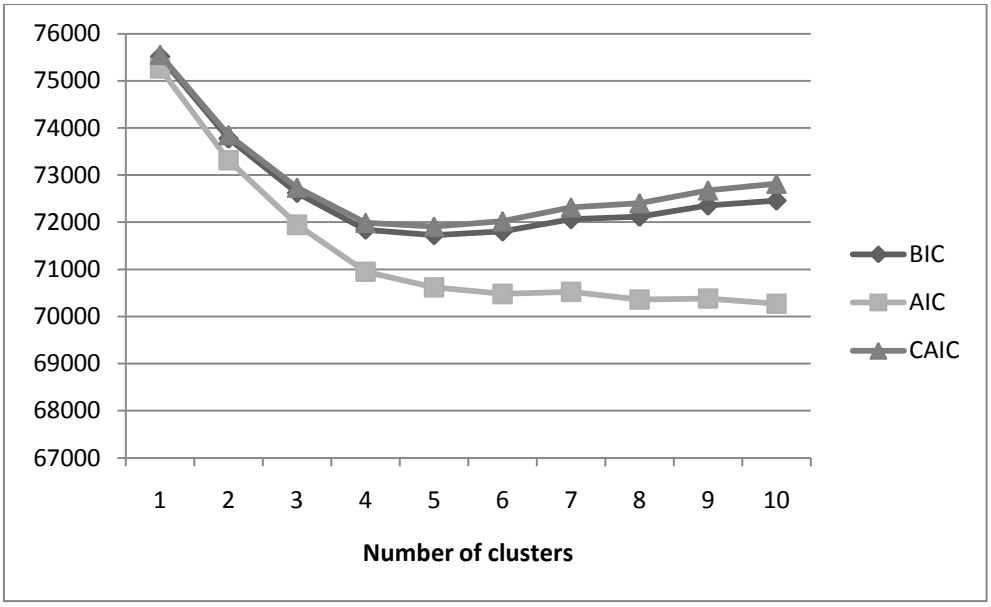

Figure 5-10: Evolution of BIC, AIC and CAIC with addition of clusters (Intersec_TR)

All road crashes at this intersection type occurred inside built-up areas. Cluster one mainly involves side collisions (82%) during the evening rush hour (30%) in autumn (30%) in good weather (97%) and road conditions (95%) and 58% of drivers failed to give priority to vehicles coming from the right. This cluster is typically distinct from the others as it is the only cluster dominated by right of way to traffic from the right violation. It is referred to as "road crashes in normal weather and road conditions with right of way violation".

Cluster two comprises road crashes that occurred in summer (29%) during precipitation (61%) on wet roads with puddles (83%). This cluster is referred to as "road crashes in summer on wet roads with puddles during precipitation".

The third cluster comprises crashes during evening rush hour (29%) mostly during the autumn season (35%) in normal weather (96%) and road conditions (89%) involving 35% of rear-end collisions. This cluster will be labeled as *"road crashes in autumn during normal weather and road conditions".*

Cluster 4 contains mostly night crashes (98%) with public lighting (94%) during the summer season (27%) with about 46% of the crashes occurring during the week-end. This cluster is labeled as "road crashes that occurred at night with public lighting".

The last cluster contains almost entirely collisions between drivers and pedestrians (99%) hence almost 100% of the victims were pedestrians during the summer (28%) and autumn seasons (26%). This cluster is called "*road crashes with collisions between drivers and pedestrians".*

Table 5-11: Crash types at Intersec_TR

Cluster	Crash type	Size(%)
1	Road crashes in normal weather and road conditions with right of way violation.	57
2	Road crashes in summer on wet roads with puddles during precipitation.	14
3	Road crashes in autumn during normal weather and road conditions.	14
4	Road crashes that occurred at night with public lighting.	8
5	Road crashes with collisions between drivers and pedestrians.	7

5.4 Discussion and Summary

5.4.1 Interpretation of results of pre-analysis using Descriptive statistics

The pre-analysis of the data using frequency tables and charts revealed some interesting information regarding road safety in Belgium. About 52% of traffic crashes at intersections in Belgium for the year 2005 occurred on intersections where a right- of-way sign B1 (yield or give way sign) or B5 (stop sign) was present (*Intersec_RS*). This can be attributed to the fact that there are many intersections of this type or simply because these road signs can be easily violated. Moreover, taking the injury level into consideration, *Intersec_RS* comes second in terms of fatal crashes with 1.18% and highest in terms of severe injuries (10.11%). Surprisingly, intersections with a traffic policeman (*Intersec_TP*) had the highest fatality rate with 1.32%. However, the number of crash victims at this intersection type made up just 0.45% of overall crash victims.
The least fatality rate was recorded at intersections were there was a right of way to traffic from the right with 0.48% (*Intrsec_TR*). However, usually, such types of intersections are common on roads where the traffic intensity is low.
Furthermore, no traffic crash with casualty was recorded at intersections with a defective three-coloured traffic lights or amber flashing light and a right of way sign B1 or B5 was

present (*Intersec_DTLRS*). That notwithstanding, it cannot be hastily concluded that such intersections are safer. This can be attributed to the existence of few intersections of this type. Furthermore, it can be due to the stochastic nature of traffic crashes. This means that, the number of crashes at a specific road intersection can be high for a given year and declines the following year without any safety intervention due to mere chance. A possible way can be to analyse the crash rate over several years to incorporate this aspect of randomness.

Regarding the type of collision, side collision between drivers (*COL_T3)* was the most common. More than half the number of crashes at intersections (57%) involved side-collision between drivers. It had the third highest fatality rate of 0.9% in terms of crash severity and is ranked fourth in terms of serious injuries and had the second highest rate of slight injuries. Side collisions, also known as *"T-bones"* (especially in the United States of America) happen when a vehicle strikes another on the side. It sometimes occurs when a driver ignores a red traffic, disregards a stop (B5) or yield (B1) sign or a right of way to traffic from the right at an intersection or another vehicle is struck by an oncoming vehicle.

Rear-end collision (also known as *rear-end shunts* especially in the United Kingdom) was second in terms of frequency with 15.46% of crashes and registered the least rate of fatal crashes (0.4%) and seriously injured casualties of 5% and the highest rate of slightly injured victims of close to 95%. This type of collision occurs when a vehicle hits the rear of the vehicle in front and could be as a result of tailgating or panic stopping. Tailgating is the practice of driving too close to the preceding vehicle without a reasonable safety distance to stop the car effectively when the brake is applied especially if the car in front stops abruptly (panic breaking). This collision type had the least possibility of victims sustaining very serious injuries as depicted by the above figures.

Head-on or frontal collision was the third most frequent with more than 14% of all collision types and had a fatality rate of 0.7% and also ranked third in both serious injuries and slight injuries with a rate of 10% and 89% respectively. As Thomas and Frampton (1999, p.1) noted, "the risk of injury in side impacts is generally higher than in frontal crashes because there is less vehicle structure within which to attenuate crash forces…". That notwithstanding, head-on collisions are still deem to be more fatal especially at high speeds than other types of collision. However, even though it generally has a lower frequency, high severity levels have been reported in some countries. For example, statistics from the United States show that in 2005, head-on crashes made up just 2% of all crashes, but accounted for 10.1% of overall fatalities (NHTSA, 2006).

Driver-pedestrian collision came in the fourth rank in terms of frequency with 5.7% of crashes at intersections and was the second highest regarding fatality rate (2%) and the most severe in terms of serious injuries with 16% and had the least share of slightly injured victims of 82% (together with collision with an obstacle). Collision with an obstacle came next with the highest fatality rate of 2.7% and just like driver-pedestrian collisions, had the highest rate of serious crashes and consequently the least rate of slight injuries. This is in accordance with the report from the European Transport Safety Council (ETSC, 1998), which states that collisions with off-road obstacles contribute between 18 and 42% of fatal road crashes in several European countries. The report further states that, they are mostly single vehicle accidents, involving young drivers, excess or inappropriate speeding, use of alcohol or drugs and driver fatigue. Among the obstacles struck by run-off vehicles are trees, barriers, poles, lamp posts and other street furniture. That notwithstanding, improvements have been made over the years in Belgium and several European countries to make roads more forgiving and self-explanatory.
A minimal proportion of other collision types (or unknown) made up a share of 1.6%.

Regarding the time of the day, most crashes took place during the evening rush hour (4-6pm) making up 26% of the casualties. This is generally the period when commuters return from work and this can be partly attributed to driver fatigue and the increased traffic flow. This is followed by the afternoon period making up about 19% of crashes and the morning period making up 17%. The least number of crashes was recorded at night with 10% of crashes at intersections.
There are no significant fluctuations of the number of crash victims that were recorded during the different seasons. At least more than 20% of crashes were recorded during each season. 28% of crashes were registered in autumn (September to November), 26% in summer (June to August), 25% in spring (March to May) and 22% in winter (December to February). It is common for many people to travel during the summer period especially for leisure trips thanks to the sunny weather. Hence, the km travelled per vehicle is also higher during this period which also increases exposure to road crashes. The slight reduction of crashes in winter can be attributed to the poor weather marked by snow and rainfall which reduces mobility. This can take the form of trip cancellation for less important trips or change of mode to public transport. Another possible explanation can be as a result of risk compensation. This means that, faced with the risky driving conditions such as snow-covered roads, drivers may tend to drive more carefully as they perceive a higher chance of getting involved in an accident.

More than 60% of road crashes at intersections occurred inside built-up areas. This is likely the case as there tends to be many intersections inside built-up areas. Finally, 24% of crashes were recorded during the week-end (Saturday and Sunday) while the rest occurred during the weekdays. Drunk, reckless driving and consequently excessive and inadequate speeding are higher during the week-end as these are usually periods for partying and feasting.

5.4.2 Summary of Crash types

The main analysis involves clustering the different data sets representing the various intersection types in Latent Gold. The final models derived at the end of the procedure each provides cluster-dependent univariate distributions for each selected variable, with each cluster depicting a specific crash type. Several variables were used to segment the crashes including the collision type, type of road user, location of the intersection (inside built-up area or not) and weather and road conditions. It should be recalled that observations are assigned to clusters based on their posterior probabilities.

The intersection type (*Traffic-Con2*) with functioning three-coloured traffic lights has six crash types. Crashes at signal-controlled intersections usually occur when road users fail to respect the traffic signal and runs into an oncoming road user. The first crash type makes up 33% and contains those crashes that occurred in normal weather and road conditions. The second cluster principally contains crashes that resulted from traffic light violation making up 18%. Moreover, rear-end collisions are likely to be high as drivers or other road users may react slowly to the changing signals, marked by abrupt braking, giving the other road users insufficient time to brake safely. This is reflected in the third cluster which makes up 16%. Road crashes that occurred at night had an equal share of 13% with crashes during precipitation as denoted by clusters four and five. Finally, driver-pedestrian collision made up 7% of crashes at this intersection type.

The intersection type with a right-of-way sign B1 (yield) or B5 (stop) present (*Intersec_RS*) has seven crash types. This intersection recorded the highest number of crashes with more than half the overall accidents at intersections. The first cluster makes up 30% and is mostly comprised of crashes in normal weather and road conditions. Crashes that resulted from right-of-way signs B1 (yield) and B5 (stop) violation are very high making up 27%. The third cluster mainly contains rear-end collisions and make up 15% and the fourth cluster which describes crashes during precipitation makes up 13%.

The remaining clusters are night crashes making up 9%, driver-pedestrian collision (4%) and crashes at intersecting regional roads for clusters five, six and seven respectively.

The fifth intersection type with a right of way to traffic from the right (*Intersec_TR*) has five crash types. The first cluster contains 57% of crashes and reflects those crashes that occurred when there was a right of way violation. This clearly indicates that traffic violation at this intersection is very high. The next cluster reflects crashes that occurred during precipitation making up 14%. Night crashes make up 8% and driver-pedestrian collisions make up 7% as reflected in clusters four and five.

Table 5-12: Summary of crash types at various intersections

Data set	Number of clusters	Entropy criterion (*I*)
Intersec_TL	6	0.84
Intersec_RS	7	0.86
Intersec_TR	5	0.88

5.4.3 Comparison of crash types at various intersections

The second part of the analysis entails making a comparison among the various crash types that were identified at the various intersections. Some crash types identified shared a lot of similarities among the different intersections. For instance, the majority of crashes were concentrated on intersecting municipal roads (or other road type except motorway or regional roads). Only the last cluster of intersection with a right-of-way sign B1 or B5 was dominated by crashes on intersecting regional roads.

Traffic violation of right of way or priority is a common cause of crashes at the various intersection types though with varying degree of intensity. At priority-ruled intersections or intersections with priority to vehicles from the right (*Intersec_TR*), it was particularly very high. In 57% of crashes at this intersection type, at least one driver failed to give way to vehicles coming from the right. At *Intersec_RS*, (intersection with right-of-way sign B1 or B5) 27% of crashes consisted of instances when a driver failed to respect a B1 (yield) or a B5 (stop) sign. Finally, at intersections with a functioning traffic light, 18% of the crashes involved instances when the driver ignored a red light.

Even though side-collisions dominated the crash types, rear-end collisions also had a significant share of crashes at Intersec_TL and Intersec_RS. Rear-end collisions made up 18% of crashes at intersections with functioning traffic lights while it was 15% at intersections with B1 or B5 signs. Fewer rear-end collisions occurred at intersections with priority to traffic from the right.

Furthermore, crashes during precipitation (rainfall, snowfall, and hailstorm) and strong winds were noticeable at the various intersection types. The percentage of crash share was virtually the same for the three intersection types under consideration. Intersec_TL and Intersec_RS had 13% each while Intersec_TR had 15%. However, for Intersec_RS and Intersec_TR, the crashes were concentrated during the summer season while that for Intersec_TL was in the winter season.

Moreover, driver-pedestrian collisions were also significant at the three types of intersection under consideration. It made up 7% of crashes at both Intersec_TL and Intersec_TR and 4% of dominant crashes at Intersec_RS.
The dominant crash types for Intersec_TR all occurred inside built-up areas. This is comprehensible as intersections with right of way to the right (uncontrolled intersections) are usually common on roads with lower speeds and lower traffic intensity which are more likely to be found in agglomerations.

Night crashes with public lighting were also dominant at the various intersection types. Intersections with a functioning traffic light had the highest crash share as depicted in cluster four with 13%. Intersections with a right-of-way sign and that with right of way to traffic from the right both made up 9% and 8% respectively.

Chapter 6: CONCLUSION AND FURTHER RESEARCH

6.1 Conclusion and Findings

Even though the rate of road traffic fatalities and injuries have been on the decline in Belgium and several developed countries, the current level still remains unacceptably high. As such, road safety remains a top priority in Belgium especially regarding the safety of vulnerable road users including pedestrians, bicyclists and motorcyclists. In this regard, the European Union encourages the dissemination of relevant road safety measures that can further improve the prevailing road safety situation in its member countries.

Research into the causes of road crashes has led to the accumulation of huge, multidimensional and usually heterogeneous data sets regarding several aspects that surround these crashes. This has been aided by the advent of high performance computing algorithms. However, traditional data analysis techniques are no longer capable of dealing effectively with such complex data. Therefore, advanced statistical techniques are needed to retrieve some useful information from this huge bulk of traffic data. This loophole has been filled with emerging techniques from the domain of data mining. The importance of cluster analysis in segmenting crash data is the focus of this report. The main objective was to identify and compare dominant crash types at various types of intersections. A statistically based clustering technique called latent class cluster analysis (LCCA) or finite mixture models was applied to crashes that occurred at intersections in Belgium using data for the year 2005. A pre-analysis was conducted using descriptive statistics to gain a prior notion about crashes at road intersections. Unlike the distance-based traditional clustering methods such as the K-means and hierarchical clustering, this novel clustering technique has not been extensively applied in the domain of traffic safety. That notwithstanding, a few studies conducted so far have yielded good results (Geurts, Wets, Brijs and Vanhoof, 2003; Depaire, Wets and Vanhoof, 2008; Ayramo et al., 2009).

The traffic crash data for Belgium in the year 2005 was segmented according to the different intersection types (six in all), and the respective crash types for three types were determined using the Latent Gold software. The main findings that were deduced from the analyses are presented below with regards to the research objectives:

The usefulness of latent class cluster analysis was clearly demonstrated in reducing the heterogeneity of crash data. By reducing the crash data into smaller clusters, it becomes easier to direct countermeasures to specific aspects of the multitude of factors that cause road crashes. However, crashes at intersections with a traffic policeman (*Intersec_TP*) and intersection with a defective traffic light and right of way to traffic from the right (*Intersec_DTLTR*) could not be effectively modeled in latent Gold due to fewer observations and little within variation. The dominant crash types at the three intersection types (*Intersec_TL, Intersec_RS and Intersec_TR*) which had sufficient observations were identified. Intersection with a functioning three-coloured traffic light (*Intersec_TL*) had six crash types, intersection with a right-of- way sign B1 or B5 (*Intersec_RS*) had seven crash types and intersections with right of way to traffic from the right (Intersec_TR) had five crash types. The crash types did not differ significantly among the intersection types and could be summarized as follows according to frequency:

- Road crashes in normal weather and road conditions (mostly involving side-collisions).
- Road crashes with traffic light/ right of way sign B1 or B5/ right of way to traffic from the right violation.
- Road crashes with rear-end collisions.
- Road crashes at night with public lighting.
- Road crashes during precipitation (and strong winds).
- Road crashes with collisions between drivers and pedestrians

These crash types shared similarities among the various intersection types but differ in intensity.

Another objective was to describe the characteristics of crashes at intersections using basic exploratory data analysis (EDA). Important information was discerned for example the collision type, time of the day or period when the accident occurred or the season among others. For instance, more than half the percentage of crashes occurred at intersections with a right-of-way sign B1 or B5. Moreover, surprisingly, intersection with a traffic police man had the highest fatality rate. However, as stated before, it had the lowest number of crashes of just 61 cases. Also, regarding the type of collision, the frequency distribution showed that close to 60% of crashes involved side-collisions and regarding the time of day, crashes occurred frequently during the evening rush hour which can be linked to the traffic intensity as well as driver fatigue.

It is clear from the analysis that, as we move from uncontrolled (priority-ruled) to controlled intersections (signalised), priority or right of way violation which is a key cause of crashes at intersections, tend to decline. Hence, intersections with functioning three-coloured traffic lights tend to be safer compared to priority rule intersections in this regard. It is therefore recommendable to install traffic lights at intersections where the traffic intensity renders it worthwhile.

6.2 Limitations and Further Research

There are certain shortcomings about the method of latent class clustering which can lead to biased results. These include:

- The occurrence of boundary solutions: These occur when estimated probabilities equal to 0 or 1, or the log-linear parameters equal to or minus (or plus) infinity. The resultant effect can be complications in the estimation algorithms, occurrence of local solutions, difficulties in computing the standard errors and the number of degrees of freedom (df) of the goodness-of-fit tests. However, these can be avoided by imposing constraints or taking into consideration other types of information on the model parameters.
- The presence of local solution or local maxima: Given that the log-likelihood function of latent class models is not always concave; the obtained solution can be the local instead of the global solution. This implies that the solution depends on the initial or starting parameter values. This problem is resolved in Latent Gold by estimating the model with different sets of random starting values. Several sets of values therefore converge to the same highest log-likelihood value, thereby reducing the possibility of a local solution. The Latent Gold programme uses 10 different sets of starting values to minimize the occurrence of local solutions.
- Local independence assumption: The fundamental assumption underlying latent class or finite mixture models is that of local independence among crash variables. It is possible to relax this assumption on the covariance matrix but this will cause parametric complexity and longer computing times.

Further research on the identification of dominant crash types can focus on standardizing relevant variables that can be used to describe crashes at intersections. This will facilitate comparison between regions or countries or from one period to another. Moreover, due to the stochastic or random nature of road crashes, it can be of importance to use data for several years which is representative of the *"true"* number of

crashes or casualties. This will also ensure that there are enough observations for each intersection type for which cluster analysis can be conducted such as the cases of intersection with a traffic policeman (Intersec_TP) and that with a defective three-coloured traffic light or amber flashing light and there was right of way to traffic from the right (Intersec_DTLTR) which did not have sufficient observations. This was also the case for intersections with a defective three coloured traffic light or amber flashing light and there was right of way sign B1 or B5 (Intersec_DTLRS) which did not have any observation at all. This will produce more realistic results as common pitfalls in dealing with crash data such as regression to the mean (RTM) and the migration of crashes can be mitigated.

In a nutshell, latent class cluster analysis (LCCA) is an effective technique which can be used to reduce the heterogeneity of road crash data. As Cameron (1992) stated, clustering techniques in general, are important tools when analysing traffic accidents as these methods are capable of identifying groups of road users, vehicle types and road segments which could be suitable targets for countermeasures.

Bibliography

Al-Balbissi, A. H., 2003. Role of Gender in Road Accidents, *Traffic Injury Prevention*, 4, pp.64-73.

Arabie, P. and Hubert, L.J., 1994. Cluster analysis in marketing research. In: Bagozzi, R.P. ed., Advanced Methods of Marketing Research. Blackwell, Oxford, pp.160-189.

Arminger, G. and Stein, P., 1997. Finite mixture of covariance structure models with regressors: loglikehood function, distance estimation, fit indices, and a complex example". *Sociological Methods and Research* 26, pp.148-182.

Atkeson, C.G., Moore, A.W. and Schaal, S., 1997. Locally Weighted Learning. *Artificial Intelligence Review*, 1-5, pp.11-73.

Ayramo, S., Pirtala, P., Kauttonen, J., Naveed, K. and Karkkainen, T., 2009. Mining Road Traffic Accidents. Report of the Department of Mathematical Information Technology, Series C(2). Software and Computational Engineering, Jyvaskyla, Finland.

Baker, D.R., Clarke, S.R. and Brandt, E.N., 2000. An Analysis of Factors Associated with Seat Belt Use: Prevention Opportunities for the Medical Community. *The Journal of the Oklahoma State Medical Association*. 93(10), pp.496-500.

Banfield, J.D. and Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian Clustering." Biometrics 49(3), pp.803-821.

Batchelor, B.G., 1978. *Pattern Recognition: Ideas in Practice*. New York: Plenum Press.

Bayam, E., Liebowitz, J. and Agresti, W., 2005. Older drivers and accidents: A meta analysis and data mining application on traffic accident data. *Expert Systems with* Applications, 29(1), pp.598-629.

Bédard, M., Guyatt, G.H., Stones, M.J. and Hirdes, J.P., 2002. The Independent Contribution of Driver, Crash and Vehicle Characteristics to Driver Fatalities. *Accident Analysis and Prevention*, 34(6), pp. 717-727.

Belgian road network. [online]: Available at: www.europe.aaroads.com. [Accessed 10 January 2011].

Bensmail, H., Celeux, G., Raftery, A.E. and Robert, C.P., 1997. Inference in model based cluster analysis. *Statistics and Computing* 7, pp.1-10.

Bird R.N., 2001. Junction Design. In: Bulton, K.J. and Hensher D.A. eds. 2001. Handbook of Transport Systems and Traffic Control. Elsevier Science Limited.

Bockenholt, U., 1993. A Latent Class Regression Approach for the Analysis of Recurrent Choices. *British Journal of Mathematical and Statistical Psychology* 46, pp.95-118.

Bradley, P.S., Mangasarian, O.L and Street, W.N., 1997. Clustering via Concave Minimization. In: Advances in Neural Information Processing Systems, 9. Mozer, M.C, Jordan, M.I. and Petsche, T, eds. Pp.368-374, Cambridge, MA: MIT Press, pp.368-374.

Brijs, T., 2002. Retail Market Basket Analysis. A Quantitative Modeling Approach. Ph.D.dissertation. University of Hasselt, Diepenbeek, Belgium.

Brussels Ring [online]: Available at: http://en.wikipedia.org/wiki/Brussels_Ring. [Accessed 10 January 2011].

Cameron, M.,1992. Accident Data Analysis to Develop Target Groups for Countermeasures: Methods and Conclusion, 1. Monash University Accident Research Centre (MUARC),

Celeux, G., Biernacki, C., and Govaert, G., 1997. Choosing Models in Model-based Clustering and Discriminant Analysis. Technical Report. Rhone-Alpes: INRIA.

Chang, L.Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety Science*, 43(8), pp.541-557.

Chang, L. and Chen, W., 2005. Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, 36, pp.365-375.

Chang,L.Y. and Wang, W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, 38(5), pp.1019-1027.

Cheeseman, P. and Stutz, J., 1995. Bayesian classification (Auto class): Theory and results. In: Advances in knowledge discovery and data mining. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds. Menlo Park: The AAAI Press.

Chen, W. and Jovanis, P., 2002. Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record*, 1717, pp.1-9.

Chengqian, M., Jingling, Y., Zhong, L. and Yue, X., 2009. Data Mining in Traffic Flow Analysis of City Tunnel, *Database Technology and Applications*, pp.199-201. First International Workshop on Database Technology and Applications, Wuhan Hubei, China.

Chimba, D. and Sando, T., 2009. The prediction of highway traffic accident injury severity with neuromorphic techniques. *Advances in Transportation*, A(19),pp.17-26.

Chong, M., Abraham, A. and Paprzycki, M., 2005. Traffic accident Analysis using Machine Paradigms. Informatica, 29, pp.89-98.

Clogg, C.C., 1981. New developments in Latent Structure Analysis." Pp. 215-246, in *Factor Analysis and Measurement in Sociological Research*, eds. D.J. Jackson, D.J and Borgotta, E.F. Beverly Hills: Sage Publications.

Clogg, C.C., 1995. Latent Class Models. pp. 311-359, In: *Handbook of statistical modeling for the social and behavioral sciences,* eds. Arminger, G., Clogg, C.C and Sobel, M.E. New York: Plenum Press.

Daniels, S., Nuyts, E. and Wets, G., 2008. The effects of roundabouts on traffic safety for bicyclists: an observational study. *Accident Analysis & Prevention* 40(2), pp.518–526.

Dayan, P., 1999. Unsupervised Learning. In: Robert, A.W. and Keil, F.C., 1999. eds. MIT Encyclopedia of the Cognitive Sciences. MIT Press.

Dean, N. and Raftery, A.E., 2010. Latent Class Variable Selection. Annals of the Institute of Statistical Mathematics, 62(1), pp.11-35.

Delen, D., Sharda, R. and Bessonovi, M., 2006. Identifying significant predictors of injury severity in traffic accident series of artificial neural networks. *Accident Analysis and Prevention*, 38, pp.434-444.

Depaire, B., Wets, G. and Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention* 40, pp.1257-1266.

Diday, E., 1974. Recent Progress in Distance and Similarity Measures in Pattern Recognition. Second Joint Conference on Pattern Recognition, pp.534-539.

Dolan, C.V. and Van der Maas, H.L.J., 1997. Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, 63, pp.227-253.

Dougherty, M., 1995. A review of neural networks applied to transport. *Transportation Research*, Part C, 3(4), pp.247-260.

Duda, R.O., Hart, P.E. and Stork, D.G., 2001.*Pattern Classification.* 2nd ed. New York: John Wiley.

Dunham, M.H., 2003. *Data mining: introductory and advanced topics*, Pearson Education Inc, Upper Saddle River, New Jersey.

Elvik, R., 2003. Effects on road safety of converting intersections to roundabouts: review of evidence from non-U.S. studies. Transportation Research Record: *Journal of the Transportation Research Board* 1847, pp.1–10.

Elvik, R., Hoye, A., Vaa, T. and Sorensen, M., 2009. *The Handbook of Road Safety Measures*, 2nd ed. Bingley: Emerald Group Publishing Limited.

Esnaf, Ş., Koldemir. B., Küçükdeniz, T. and Akten N. 2008. Fuzzy Cluster Analysis of Shipping Accidents in the Bosporus. *European Journal of Navigation*, 6(3), pp.32-41

European Environment Agency (EEA), 1998. Europe's Environment: The Second Assessment. Office for Official Publications of the European Communities. Luxembourg.

European Road Safety Observatory (ERSO), 2006. Traffic safety Basic facts: Junctions. [online] Available at: http://ec.europa.eu/transport/roadsafety_library/care/doc/safetynet/2006/bfs2006_sn-ntua-1-3-junctions.pdf. [Accessed 20 November 2010].

European Transport Safety Council (ETSC), 1998. Forgiving Roadsides. Brussels.

European Transport Safety Council (ETSC), 2010. Road Safety Target in Sight: Making up for Lost Time, ETSC 4th Road Safety PIN Report, Brussels.

European Transport Safety Council (ETSC), 2008. ShLOW! Show me How Slow: Reducing Excessive and Inappropriate Speed Now: a Toolkit.

European Transport Safety Council (ETSC), 2009. 2010 on the horizon, 3rd Road Safety PIN Report, Brussels.

Everitt, B.S., 1988. A Finite Mixture Model for the Clustering of Mixed-mode data. *Statistics and Probability Letters* 6, pp.305-309.

Everitt, B.S., 1993. *Cluster Analysis*. London: Edward Arnold.

Everitt, B., Landau, S. and Leese, M., 2001. *Cluster Analysis*. 4th ed. London: Arnold.

Fraley, C. and Raftery, A.E., 1998a. How many clusters? Which clustering method? – *Answers via model-based cluster analysis*. Department of Statistics, University of Washington: Technical Report no. 329.

Fraley, C. and Raftery, A.E., 1998b. MCLUST: Software for Model-based Cluster Analysis. *Journal of Classification*, 16, pp.297-306.

Fraley, C. and Raftery, A. E., 2002. Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association*, 97, pp.611–631.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C., 1992. Knowledge Discovery in Databases: An Overview. Al Magazine, Fall 1992, pp. 213-228.

FPS-Economy, SMEs, Self-employed and Energy- Statistics Belgium., 2009. Key figures: Belgium and the European Union. [online] Available at: http://statbel.fgov.be/en/binaries/key_figures_2009_tcm327-109303.pdf. [Accessed 5 January 2011].

Garder, P., 1998. The Modern Roundabouts: The Sensible Alternative for Maine. Maine Department of Transportation, Bureau of Planning, Research and Community Services, Transportation Research Division.

Geurts, K., Thomas,I. and Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. *Accident Analysis and Prevention*, 37(4), pp.787-799.

Geurts, K., Wets, G., Brijs, T. and Vanhoof, K., 2003a. Profiling High Frequency Accident Locations Using Association Rules. Proceedings of the 82nd Annual Meeting of the Transport Research Board, January 12-16 2003, Washington D.C, USA.

Geurts, K., Wets, G., Brijs, T. and Vanhoof, K., 2003b. Clustering and Profiling of Traffic Roads by Means of Accident data. In Proceedings of the European Transport Conference. October 8-10 2003, Strasbourgh, France.

Golob, T.F and Recker, W.W., 2003. A method for relating type of crash to traffic flow characteristics on urban freeways. California PATH program, Institute of Transportation Studies, University of California, Berkeley.

Goodman, L. A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), pp.215–231.

Gordon, A.D., 1994. Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*. 18, pp.561-581.

Graepel, T., 1998. Statistical Physics of Clustering Algorithms. Technical Report 171822, FB Physik, Institut fur Theoretische Physic.

Green, P.E., 2004. Practice makes perfect. *Marketing Research,* 16(2), pp.8-14.

Guha, S., Rastogi, R. and Shim, K., 2001. CURE: An Efficient Clustering Algorithm for Large Databases. SIGMOD, Seattle, WA-USA.

Hagenaars, J.A. and McCutcheon, A.L. 2002. Applied latent class analysis. Cambridge: Cambridge University Press.

Han, J. and Kamber, M., 2000. *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann Publishers.

Harkey, D., 1995. Intersection Geometric Design Issues for Older Drivers and Pedestrians. *Compendium of Technical Papers*, Institute of Transportation Engineers 65th Annual Meeting, August 5-8, Denver, Colorado.

Hartigan, J. and Wong, M., 1979. A k-means Clustering Algorithm. *Journal of Applied Statistics*, 28, pp.100-108.

Hoornaert, B., 2010. Social Costs of Road Accidents in Belgium, European Road Safety Day, Brussels, 13th October 2010, Belgium. Federal Planning Bureau.

Hunt, L. and Jorgensen, M., 1999. Mixture Model Clustering using the MULTIMIX Program. *Australian and New Zealand Journal of Statistics,* 41, pp.153-172.

Jain, A.K., Murty, M. A. and Flyn, P., 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3), pp.264-323.[online] Available

Jedidi, K., Jagpal, H.S. and DeSarbo, W.S., 1997. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science* 16, pp.39-59.

Jiang, D., Tang, C. and Zhang, A., 2004. Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering (TKDE), 16(11), pp.1370-1386.

Jorgensen, M., and Hunt, L., 1996. Mixture model clustering of data sets with categorical and continuous variables. In: Proceedings of the Conference ISIS 1996, Australia. pp.375-384.

Kaufman, L. and Rousseeuw, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley.

Kavsek, B., Lavrac, N. and Jovanoski,V., 2006. Apriori-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence, 20(7), pp.543-583.

Kim, K., Nitz, L., Richardson, J. and Li, L., 1995. Personal and Behavioural Predictors of Automobile Crash and Injury Severity. *Accident Analysis and Prevention*, 27(24), pp.469-481.

Kleinemas, U., and Rudinger, G., 2010. Profiles of elderly drivers with accidents - An in-depth study, and its implications for intervention. Paper delivered at the 12th International Conference on Mobility and Transport for elderly and disabled persons (TRANSED 2010), 2-4 June, 2010, Hong Kong.

Kruskal, J.B., 1964. Multidimensional Scaling by Optimizing Goodness-of-fit to a Non-metric Hypothsesis. *Psychometrica,* 29(1): pp.1-27

Kulmala, R., 1994. Measuring the safety effect of road measures at junctions. Technical Research Centre of Finland, Espoo. *Accident Analysis and Prevention*, 26(6), pp.781-794.

Lance, G.N. and Williams, W.T., 1967. Mixed-data classificatory programmes I. Agglomerative systems. *Australian Computer Journal*, 1, pp.15-20.

Lawrence C.J. and Krzanowski, W.J., 1996. Mixture separation for mixed-mode data. *Statistics and Computing* 6, pp.85-92.

Lazarsfeld, P.F., 1950a. The logical and mathematical foundations of latent structure analysis. In: S. A. Stouffer ed., *Measurement and prediction, the American soldier*: *studies in social psychology in World War* II, Vol.4 Chap. 10, pp. 362–412. Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., 1950b. The interpretation and computation of some latent structures. In S. A. Stouffer ed., *Measurement and prediction, the American soldier: studies in social psychology in World War II*, Vol. 4, Chap. 11, pp. 413–472. Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. and Henry, N. W., 1968. Latent structure analysis. Boston: Houghton Mifflin.

Macqueen, J.B., 1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp.1-28.

Maycock, G. and Hall, R.D., 1984. Crashes at four-arm roundabouts. TRRL Laboratory Report LR 1120. Transport and Road Research Laboratory, Crowthorne.

McCutcheon, A.L., 1987. *Latent Class Analysis*. Newbury Park, C.A: Sage Publications Inc.

McLachlan, G.J., and Basford, K.E., 1988. Mixture Models: Inference and Application to Clustering. New York: Marcel Dekker.

McLachlan, G.J and Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley series in probability and statistics. New York: John Wiley and Sons Inc.

McLachlan, G.J and Peel, D., 2000. *Finite Mixture Models*. New York: Wiley.

McLachlan, G.J. and Peel, D., 1996. An algorithm for unsupervised learning via normal mixture models. In: Information, *Statistics and Induction in Science*, Dowe, D.L, Korb, K.B. and Oliver, J.J. eds. Singapore: World Scientific Publishing, Pp. 354-363.

McLachlan, G.J., Peel, D., Basford, K.E. and Adams, P., 1999. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software* 4(2).

Michalski, R.S., Stepp, R.E. and Diddy, E. 1981. A Recent Advance in Data Analysis: Clustering Objects into Classes Characterised by Conjunctive Concepts: *Progress in Pattern Recognition*. In: Laveen, N.K. and Rosenfield, A. eds. New York: North Holland, pp. 33-56.

Moghaddam, F.R., Afandizadeh, S. and Ziyadi, M., 2011. Prediction of accident severity using artificial neural networks. *International Journal of Civil Engineering*, 9, pp.41-49.

Moustaki, I. 1996. A latent trait and a latent class model for mixed observed variables." *The British Journal of Mathematical and Statistical Psychology*, 49(2), pp.313-334.

Moustaki, I and Papageorgiou, I., 2005. Latent class models for mixed outcomes with applications in Archaeometry. *Computational Statistics and Data Analysis*, 48(3), pp.659-675.

Mussone, L., Ferrari, A. and Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention*, 31(6), pp.705-718.

Muthen, B., and Muthen, L., 1998. Mplus: User's manual. Los Angeles: Muthen and Muthen.

Nadler, M. and Smith, E.P., 1993. Pattern Recognition Engineering. New York: John Wiley and Sons Inc.

National Highway Traffic Safety Administration (NHTSA)., 2006. Traffic Safety Facts 2006. NHTSA's National Centre for Statistics and Analysis, Washington, DC.

Organisation for Economic Co-operation and Development / International Transport Forum (OECD/ITF)., 2006. Young drivers: the road to safety. Paris.

Ossenbruggen, P.J., Pendharkar, J. and Ivan, J., 2001. Roadway safety in rural and small urbanized areas. *Accident Analysis and Prevention*.33 (4), pp. 485-498.

Ossiander, E.M. and Cummings, P., 2002. Freeway Speed Limits and Traffic Fatalities in Washington State. *Accident Analysis and Prevention*, 34, Pp. 13-18.

Oxley, J., Corben, B., Fildes, B., O'Hare, M., and Rothengatter, T., 2004. Older vulnerable road users: measures to reduce crash and injury risk. Monash University Accident Research Centre (MUARC), Melbourne, Victoria.

Pedrycz, W., 2005. *Knowledge-Based Clustering: From Data to Information Granules*. New Jersey: John Wiley and Sons.

Postorino M.N., Sarnè, G.M.L. 2001. Cluster Analysis for Road Accident Investigations. In: Proceedings of Urban Transport 2001, Wessex Institute of Technology (WIT) Southampton.

Prato, C.G, Bekhor, S., Galtzur, A., Mahalel, D. and Prashker, J., 2010. Exploring the Potential of Data Mining Techniques for the Analysis of Accident Patterns. 12th World Conference on Transport Research Society (WCTR), July 11-15, Lisbon, Portugal.

Räsänen, M. and Summala, H., 1998. Attention and Expectation Problems in Bicycle-car Collisions: An in-depth study. *Accident Analysis and Prevention*, 30(5), 657-666.

Roh, J.W., Bessler, D.A. and Gilbert, R.F., 1999. Traffic fatalities, Peltzman's model and directed graphs. *Accident Analysis and Prevention*, 31(1-2), pp. 55-61.

Romesburg, H. C., 2004. Cluster Analysis for Researchers. Morrisville, N.C: Lulu Press.

Savolainen, P.T., Mannering, F.L., Lord, D. and Qudus, M.A., 2011. The Statistical Analysis of Highway Crash-Injury Severities: A review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*.

Schoon, C.C. and Van Minnen, J., 1993. Accidents on Roundabouts: II. Second study into the road hazard presented by roundabouts, particularly with regard to cyclists and moped riders. R-93-16. The Netherlands: SWOV Institute for Road Safety Research.

Thomas, P. and Frampton, R., 1999. Injury Patterns in Side-collisions- A New Look with Reference to Current Test Methods and Injury Criteria. Vehicle Research Centre, Loughborough University, UK.

Tryon, R.C., 1939. Cluster Analysis. Ann Arbor, Michigan: Edwards Brothers.

Tseng, W.S., Nguyen, H., Liebowitz, J. and Agresti, W., 2005. Distractions and motor vehicle accidents: data mining application on fatality analysis reporting system (FARS) data files. *Industrial Management and Data Systems*, 109(9), pp.1188-1205.

Vermunt, J.K., 1997. LEM: A general program for the analysis of categorical data. User's manual. Tilburg University, the Netherlands.

Vermunt, J.K. and Magidson, J., 2000. *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K. and Magidson, J., 2002. Latent Class Cluster Analysis. In: Hagenaars, J.A., McCutcheon, A.L., eds., Applied Latent Class analysis. Cambridge: Cambridge University Press, UK, pp. 89-106.

Wedel, M., DeSarbo, W.S., Bult, J.R and Ramaswamy, V., 1993. A latent class Poisson regression model for heterogeneous count data with an application to direct mail. *Journal of Applied Econometrics*, 8, pp.329-350.

Wegen-Routes.be. [online]. www.belgianroads.tk. [Accessed 10 January 2011].

Weijmars, W. and Van Berkum, E., 2005. Analyzing highway flow patterns using cluster analysis. In: Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, September 13-16 2005. Vienna, Austria.

Wettschereck, D., Aha, D.W and Mohri, T., 1997. A Review of Comparative Evaluation of Feature Weighting Methods for Lazy Learning Algorithms. Artificial Intelligence Review, 11(1-5)

Wilson, R.A & Keil, F., eds., 1999. *The MIT Encyclopedia of the Cognitive Sciences (MITECS).* Massachusetts: MIT Press.

Wolfe, J.H., 1970. Pattern Clustering by Multivariate Cluster Analysis. Multivariate Behavioural Research, 5, pp.329-350.

World Health Organization (WHO)., 2004. World Report on Road Traffic Injury Prevention. Geneva.

Xu, R. and Wunsch II, D., 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), pp.645-678.

Yamamoto, T., Kitamu, R. and Fujii, J., 2002. Drivers' route choice behavior analysis by Data mining algorithms. *Transportation Research Record*, 1807, pp.59-66.

Yang, H., Kitamura, R., Jovavis, P.P., Vaughn, K.M. and Abdel-Aty, M., 1993. Exploration of route choice behavior with advanced traveler information using neural network concepts. *Transportation*, 20(2), pp.199-223.

Yung, Y.F., 1997. Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, pp.297-330.

Zadeh, L.A., 1965. Fuzzy Sets. Information and Control, 8, pp.338-353.

Zaiane, O.R., 1999. CMPUT 690: Principles of Knowledge Discovery in Databases. University of Alberta, unpublished.

Zeitouni, K. and Chelghoum, N., 2001. Spatial decision tree: application to traffic risk analysis. Proceedings of the ACS/IEE International Conference on Computer Systems and Applications. Beirut, Lebanon.

Appendices

Appendix 1: SAS Codes

/* Creating a library named JOHN to store the data */

```
LIBNAME JOHN "F:\DATA";
```

/* Importing the *Accidents* sheet into SAS 9.2 */

```
PROC IMPORT OUT= JOHN.accidents
            DATAFILE= "F:\DATA\accident data 2005.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="accidents$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
```

/* Selecting crashes that occurred at intersections only using SQL*/

```
PROC SQL;
  CREATE TABLE CRASH AS
  SELECT *
  FROM JOHN.accidents
  WHERE r4 = 1;
QUIT;
```

/* Adjusting the traffic control variable */

```
  IF r7_1 = "1" THEN traffic_control = "1";
  IF r7_1 = "2" THEN traffic_control = "2";
  IF r7_1 = "3" AND r7_2 = "4" THEN traffic_control = "6";
  IF r7_1 = "3" AND r7_2 = "5" THEN traffic_control = "7";
  IF r7_1 = "4" THEN traffic_control = "4";
  IF r7_1 = "5" THEN traffic_control = "5";
```

/* Creating the variable *SEASON* from months */

```
IF r3_2 IN (1 2 12)    THEN season="Winter";
  IF r3_2 IN (3 4 5)   THEN season="Spring";
  IF r3_2 IN (6 7 8)   THEN season="Summer";
  IF r3_2 IN (9 10 11) THEN season="Autumn";
```

/* Splitting hours into specific periods */

```
  IF r3_4 IN (6 7 8 9)           THEN Hour="Morning";
  IF r3_4 IN (10 11 12)          THEN Hour="Late morning";
  IF r3_4 IN (13 14 15)          THEN Hour="Afternoon";
  IF r3_4 IN (16 17 18)          THEN Hour="Evening rush";
  IF r3_4 IN (19 18 20 21)       THEN Hour="Evening";
  IF r3_4 IN (0 1 2 3 4 5 22 23) THEN Hour="Night";
```

```
/* Splitting the data based on intersection type */

DATA Intersec_TP;
  SET CRASH;
  WHERE traffic_control = "1";
 RUN;

DATA Intersec_TL;
  SET CRASH;
  WHERE traffic_control = "2";
 RUN;

 DATA Intersec_RS;
  SET CRASH;
  WHERE traffic_control = "4";
 RUN;

 DATA Intersec_TR;
  SET CRASH;
  WHERE traffic_control = "5";
 RUN;

DATA Intersec_DTLRS;
  SET CRASH;
  WHERE traffic_control = "6"


 DATA Intersec_DTLTR;
  SET CRASH;
  WHERE traffic_control = "7";
 RUN;
```

Appendix 2: Latent Gold output

2.1: Intersection with Functioning Three-coloured Traffic Lights (Intersec_TL)

Models	BIC	AIC	CAIC	Npar
1	67968	67731	68008	40
2	66808	66370	66882	74
3	65651	65012	65759	108
4	65223	64383	65365	142
5	64979	63938	65155	176
6	64514	63271	64724	210
7	64596	63152	64840	244
8	64681	63036	64958	278
9	64857	63011	65169	312
10	64944	62896	65290	346

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Cluster Size	0.3275	0.1823	0.1555	0.1343	0.1284	0.0721
ROAD_T1						
1	0	0.0332	0.0372	0.0281	0.006	0
2	0.0374	0.2353	0.242	0.1257	0.1129	0.0455
3	**0.9626**	**0.7315**	**0.7208**	**0.8462**	**0.881**	**0.9544**
ROAD_T2						
	0	0.0061	0.0047	0.008	0	0
1	0.0045	0.0426	0.0159	0.0125	0.0204	0.0051
2	0.0152	0.0955	0.0707	0.0337	0.0393	0.0202
3	**0.9803**	**0.8558**	**0.9088**	**0.9458**	**0.9403**	**0.9747**
LIGHT_C						
1	**0.8207**	**0.7613**	**0.7758**	0.0007	**0.4805**	**0.7248**
2	0.0339	0.062	0.0524	0.0288	0.0949	0.0404
3	0.1428	0.1699	0.1643	**0.9541**	0.3993	0.2298
4	0.0015	0.0068	0.0075	0.0164	0	0
9	0.0012	0	0	0	0.0254	0.0051
FATS						
	0.0075	0.0223	0	0.0322	0.009	0.0152
SER_INJ						
	0.077	0.1632	0.0845	0.1758	0.102	0.1568
SLI_INJ						
	1.2532	1.4717	1.5254	1.4033	1.3501	0.8961
HOUR						
	0.0004	0.0004	0.0005	0	0.0003	0.0003
After	0.2034	0.2017	0.2203	0	0.1742	0.1493
EvenR	**0.2559**	**0.2549**	**0.2656**	0	**0.2368**	0.218
Evenx	0.1685	0.1686	0.1676	0	0.1685	0.1666
L_mor	0.1741	0.1749	0.1659	0	0.1882	0.1998
Mornx	0.1853	0.1869	0.1692	0.0063	0.2164	**0.2468**
Night	0.0124	0.0125	0.0108	**0.9937**	0.0156	0.0191
SEASON						
Autumn	**0.3145**	**0.2991**	0.2587	0.2478	0.2239	**0.265**
Spring	0.2467	0.2434	0.2327	0.2292	0.2208	0.2346
Summer	0.2261	0.2315	0.2446	0.2478	0.2544	0.2427
Winter	0.2128	0.2261	**0.264**	**0.2751**	**0.3009**	0.2577
PEDESTRIAN						
0	**0.9968**	**0.9977**	**1**	**0.9974**	**0.9997**	0.0008
1	0.0032	0.0023	0	0.0026	0.0003	**0.9992**
MOP_C						
0	**0.8688**	**0.939**	**0.9833**	**0.9435**	**0.9251**	**0.9797**
1	0.1312	0.061	0.0167	0.0565	0.0749	0.0203
WEATHER						

1	**0.969**	**0.9679**	**0.9837**	**0.7635**	0.1837	**0.6924**
2	0.0011	0.0033	0.0001	0.1751	**0.6913**	0.2521
3	0.0006	0.0195	0.0084	0.0082	0.0086	0
4	0.0293	0.0093	0.0078	0.0532	0.1164	0.0556
ROAD_C						
1	**0.9266**	**0.8649**	**0.8223**	**0.6504**	0.0006	**0.6166**
2	0.0656	0.1351	0.1777	0.319	**0.8926**	0.3531
3	0	0	0	0.0158	0.0345	0.0101
4	0.0022	0	0	0	0	0
5	0.0056	0	0	0.0149	0.0723	0.0202
BEHAV1						
0	**0.6012**	0.2717	**0.9995**	**0.6704**	**0.6521**	**0.6463**
1	0.3988	**0.7283**	0.0005	0.3296	0.3479	0.3537
SL_D						
0	**0.8816**	0.4702	**0.53**	**0.7385**	**0.6742**	**0.8583**
1	0.1184	**0.5298**	0.47	0.2615	0.3258	0.1417
WEEKEND						
0	**0.8035**	**0.8027**	**0.7316**	**0.5394**	**0.7624**	**0.8229**
1	0.1965	0.1973	0.2684	0.4606	0.2376	0.1771
COL_TYPE						
	0.0098	0	0.0439	0.0132	0.0129	0
1	0.2181	0.2435	0.0528	0.1545	0.1653	0.0002
2	0.132	0.0306	**0.8243**	0.2407	0.2894	0.005
3	**0.5957**	**0.7187**	0.0018	**0.4096**	**0.4546**	0.0154
4	0	0	0	0	0	**0.9793**
5	0.0272	0	0.0601	0.1716	0.0524	0
6	0.0171	0.0072	0.017	0.0103	0.0252	0
BU_A						
0	**0.9205**	0.1176	0.3065	**0.5915**	**0.5809**	**0.9087**
1	0.0795	**0.8824**	**0.6935**	0.4085	0.4191	0.0913

2.2: Intersection with the Right of Way sign B1 or B5 Present (Intersec_RS)

Models	BIC	AIC	CAIC	Npar
1	164630	164356	164670	40
2	162180	161666	162255	75
3	160381	159628	160491	110
4	159183	158190	159328	145
5	157119	155886	157299	180

6	156402	154929	156617	215
7	156263	154551	156513	250
8	156445	154493	156730	285
9	156258	154066	156578	320
10	156260	153828	156615	355

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
Cluster Size	0.3038	0.2714	0.1476	0.1294	0.0872	0.0356	0.025
ROAD_T1							
1	0	0	0.0053	0.0014	0.0101	0.004	0.4079
2	0.0479	0.1364	0.1024	0.0933	0.1016	0.0444	**0.5868**
3	**0.9521**	**0.8636**	**0.8923**	**0.9053**	**0.8883**	**0.9515**	0.0054
ROAD_T2							
	0	0.0005	0	0	0.0051	0	0.0913
1	0.0037	0.0417	0.0135	0.0113	0.0359	0.004	0.3458
2	0.0027	0.0314	0.0234	0.0342	0.0377	0.0121	**0.4019**
3	**0.9936**	**0.9264**	**0.9631**	**0.9544**	**0.9213**	**0.9838**	0.1611
LIGHT_C							
1	**0.85**	**0.8702**	**0.8437**	**0.6217**	0.0012	**0.6896**	**0.7239**
2	0.0504	0.0442	0.0406	0.102	0.0395	0.0646	0.0627
3	0.0974	0.0818	0.1101	0.2565	**0.9298**	0.2337	0.1924
4	0.0017	0.0038	0.0056	0.0099	0.0279	0.0121	0.021
9	0.0006	0	0	0.0098	0.0016	0	0
FATS							
	0.004	0.028	0.0078	0.0078	0.0405	0.0483	0.0061
SER_INJ							
	0.0944	0.2072	0.0967	0.107	0.1895	0.1926	0.1303
SLI_INJ							
	1.1456	1.231	1.3015	1.3438	1.2343	0.8816	1.5032
HOUR							
After	0.2074	0.222	0.223	0.1725	0	0.1542	0.1949
EvenR	**0.2951**	**0.3044**	**0.305**	**0.2694**	0	0.2539	0.2865
Evenx	0.1201	0.1194	0.1193	0.1204	0	0.1196	0.1205
L_mor	0.1757	0.1683	0.1678	0.1934	0	0.2024	0.1821
Mornx	0.1957	0.1806	0.1796	0.2365	0.0231	0.2609	0.2096
Night	0.0059	0.0053	0.0052	0.0078	**0.9769**	0.0091	0.0065
SEASON							
Autumn	**0.3086**	**0.2902**	**0.2782**	0.1972	**0.2695**	0.2406	0.2207
Spring	0.2614	0.2574	0.2544	0.2258	0.2521	0.2432	0.2358
Summer	0.2492	0.257	0.262	**0.2912**	0.2655	**0.2767**	**0.2837**
Winter	0.1808	0.1953	0.2053	0.2858	0.2129	0.2395	0.2598
PEDESTRRIAN							

0	**0.9995**	**1**	**0.9987**	**0.9968**	**0.9984**	0.0006	**1**
1	0.0005	0	0.0013	0.0032	0.0016	**0.9994**	0
MOP_C							
0	**0.8122**	**0.892**	**0.9255**	**0.8874**	**0.926**	**0.9636**	**0.9997**
1	0.1878	0.108	0.0745	0.1126	0.074	0.0364	0.0003
WEATHER							
1	**0.9708**	**0.9669**	**0.952**	0.256	**0.8053**	**0.8332**	**0.8399**
2	0.0012	0.0008	0.0023	**0.6295**	0.1138	0.1264	0.0839
3	0.0045	0.0051	0.0138	0.0255	0.034	0.004	0.0172
4	0.0235	0.0271	0.032	0.0889	0.047	0.0364	0.059
ROAD_C							
1	**0.9283**	**0.9203**	**0.8713**	0.0045	**0.6941**	**0.7614**	**0.6916**
2	0.0685	0.0774	0.1045	**0.8889**	0.2772	0.2023	0.3001
3	0.0006	0.0007	0.0109	0.0475	0.0193	0.0242	0.0034
4	0	0.0016	0.0107	0	0	0	0
5	0.0026	0	0.0026	0.0591	0.0094	0.0121	0.0049
BEHAV1							
0	0.4886	0.3661	**0.9997**	**0.5704**	**0.6903**	**0.6819**	**0.7072**
1	**0.5114**	**0.6339**	0.0003	0.4296	0.3097	0.3181	0.2928
SL_D							
0	**0.9172**	**0.5557**	**0.7479**	**0.7203**	**0.7383**	**0.8697**	**0.5712**
1	0.0828	0.4443	0.2521	0.2797	0.2617	0.1303	0.4288
WEEKEND							
0	**0.812**	**0.777**	**0.7052**	**0.7991**	**0.5268**	**0.8103**	**0.7338**
1	0.188	0.223	0.2948	0.2009	0.4732	0.1897	0.2662
COL_TYPE							
	0.0053	0	0.0673	0.0219	0.0329	0	0.0221
1	0.1322	0.134	0.1104	0.1386	0.117	0.003	0.088
2	**0.0503**	0.0173	**0.6234**	0.1766	0.097	0.0121	0.2952
3	0.793	**0.8347**	0.0083	**0.6174**	**0.3838**	0.0012	**0.4441**
4	0	0	0	0	0	**0.9721**	0
5	0	0.0017	0.1589	0.0345	0.3594	0.0116	0.1377
6	0.0192	0.0124	0.0318	0.011	0.0098	0	0.0129
BU_A							
	0	0	0	0.0011	0	0	0
0	**0.9958**	0.0723	0.4119	0.4837	**0.5363**	**0.8701**	0.0914
1	0.0042	**0.9277**	**0.5881**	**0.5152**	0.4637	0.1299	**0.9086**

2.3: Intersection with Right of Way to Traffic from the Right (Intersec_TR)

Models	BIC	AIC	CAIC	Npar
1	75513	75266	75553	40
2	73775	73312	73850	75
3	72624	71946	72734	110
4	71843	70949	71988	145
5	**71728**	**70618**	**71908**	**180**
6	71806	70479	72021	215
7	72066	70523	72316	250
8	72117	70358	72402	285
9	72355	70381	72675	320
10	72461	70271	72816	355

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster Size	0.5729	0.1371	0.1359	0.0795	0.0746
ROAD_T1					
1	0.0015	0.0021	0.0021	0.0036	0
2	0.0091	0.0181	0.008	0.0107	0.0114
3	**0.9894**	**0.9798**	**0.9899**	**0.9857**	**0.9886**
ROAD_T2					
	0.0005	0	0	0	0
1	0.0038	0.0136	0.0036	0.0071	0
2	0.0015	0	0.0041	0.0036	0
3	**0.9942**	**0.9864**	**0.9923**	**0.9893**	**1**
LIGHT_C					
1	**0.8781**	**0.6539**	**0.8338**	0.0006	**0.7448**
2	0.0351	0.0842	0.0643	0.0048	0.0532
3	0.0837	0.2264	0.0944	**0.9479**	0.1944
4	0.0031	0.0168	0.0075	0.0467	0
9	0	0.0186	0	0	0.0076
FATS					
	0.0052	0.0021	0.0113	0.0106	0.0076
SER_INJ					
	0.0866	0.0828	0.1217	0.1151	0.1719
SLI_INJ					
	1.208	1.1987	1.1676	1.3432	0.9075
HOUR					
After	0.2376	0.203	0.2152	0	0.1642
EvenR	**0.2992**	**0.2786**	**0.2864**	0	**0.2501**
Evenx	0.1232	0.125	0.1245	0	0.1245
L_mor	0.1728	0.1911	0.1846	0	0.2113
Mornx	0.1585	0.191	0.1789	0.0236	0.2343

Night	0.0087	0.0114	0.0104	**0.9764**	0.0156
SEASON					
Autumn	**0.3034**	0.2179	**0.3462**	0.2703	0.2628
Spring	0.2466	0.2218	0.2537	0.2388	0.2368
Summer	0.2607	**0.2935**	0.2418	**0.2744**	**0.2773**
Winter	0.1892	0.2668	0.1583	0.2165	0.2231
PEDESTRI					
0	**0.9986**	**0.9938**	**0.9958**	**0.9965**	0.0007
1	0.0014	0.0062	0.0042	0.0035	**0.9993**
MOP_C					
0	**0.8466**	**0.8644**	**0.8189**	**0.9058**	**0.9543**
1	0.1534	0.1356	0.1811	0.0942	0.0457
WEATHER					
1	**0.9711**	0.249	**0.9584**	**0.7819**	**0.8098**
2	0	**0.6087**	0.0001	0.158	0.1104
3	0.0031	0.0138	0.0023	0.0072	0.0038
4	0.0258	0.1285	0.0392	0.0529	0.076
ROAD_C					
1	**0.9448**	0.0202	**0.8927**	**0.6834**	**0.7111**
2	0.0539	**0.8345**	0.073	0.2859	0.2282
3	0	0.0552	0.005	0.0172	0.0076
4	0.0013	0.0055	0.0162	0.0036	0
5	0	0.0846	0.0131	0.01	0.0532
BEHAV1					
0	0.419	**0.5524**	**0.9974**	**0.6462**	**0.7634**
1	**0.581**	0.4476	0.0026	0.3538	0.2366
SL_D					
0	**0.9085**	**0.9171**	**0.935**	**0.9085**	**0.9734**
1	0.0915	0.0829	0.065	0.0915	0.0266
WEEKEND					
0	**0.7831**	**0.8322**	**0.7176**	**0.5441**	**0.787**
1	0.2169	0.1678	0.2824	0.4559	0.213
COL_TYPE					
	0.0021	0.0166	0.0831	0.0419	0
1	0.1409	0.1704	0.3167	0.1341	0.0006
2	0.0226	0.1003	**0.3494**	0.0447	0.0024
3	**0.8241**	**0.6743**	0.0058	**0.5106**	0.0022
4	0	0	0	0.0001	**0.9948**
5	0.0014	0.0166	0.191	0.2654	0
6	0.0088	0.0218	0.0539	0.0032	0
BU_A					
	0	0.0021	0	0	0
0	**0.8301**	**0.8334**	**0.7981**	**0.8466**	**0.9885**
1	0.1699	0.1645	0.2019	0.1534	0.0115

Appendix 3: Additional Descriptive Statistics for Intersec_TP

Table 3.1a. Frequency distribution of COL_T

COL_T	Frequency	Percentage (%)
COL_T1	9	46.67
COL_T2	12	20
COL_T3	28	15
COL_T4	8	13.33
COL_T5	3	5

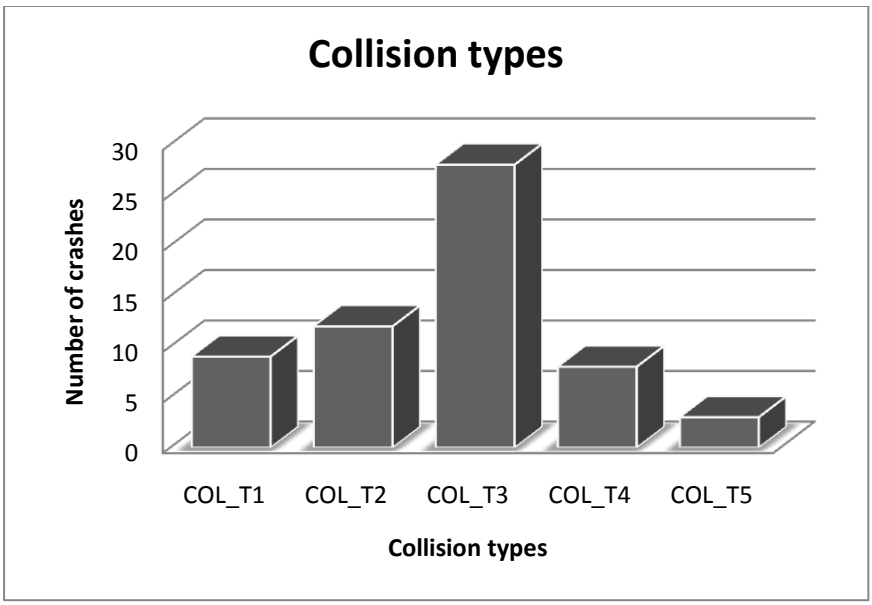

Figure 3.1b: Collision types

Table 3.2a: Frequency distribution of *HOUR*

Hour	Frequency	Percentage (%)
Mornx	9	14.75
L.Mor	9	14.75
After	16	26.23
EvenR	15	24.59
Evenx	8	13.11
Night	4	6.56

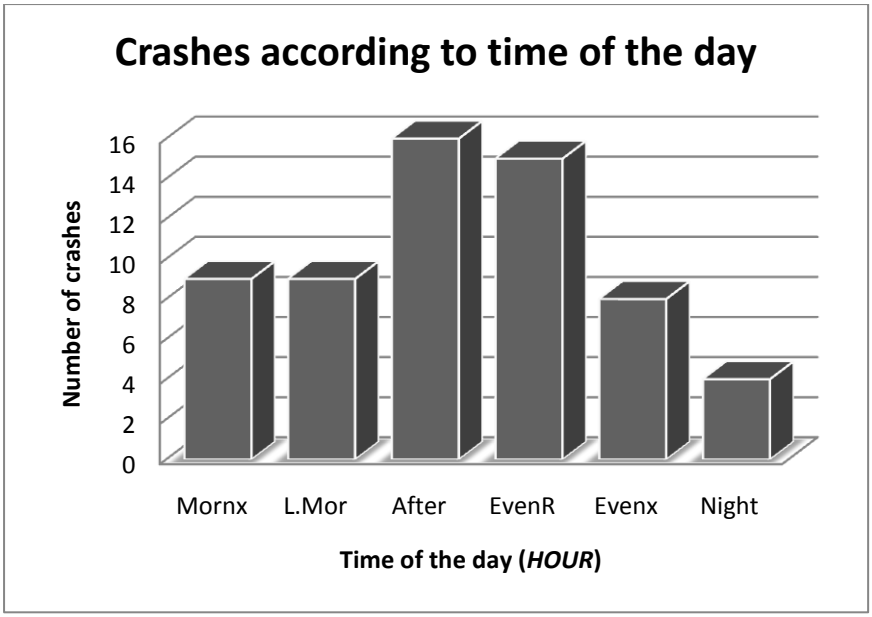

Figure 3.2b: Crashes according to time of the day (*HOUR*)

Figure 3.3a: Crashes during the week-end

Week end	Frequency	Percentage (%)
Yes	18	29.51
No	43	70.49

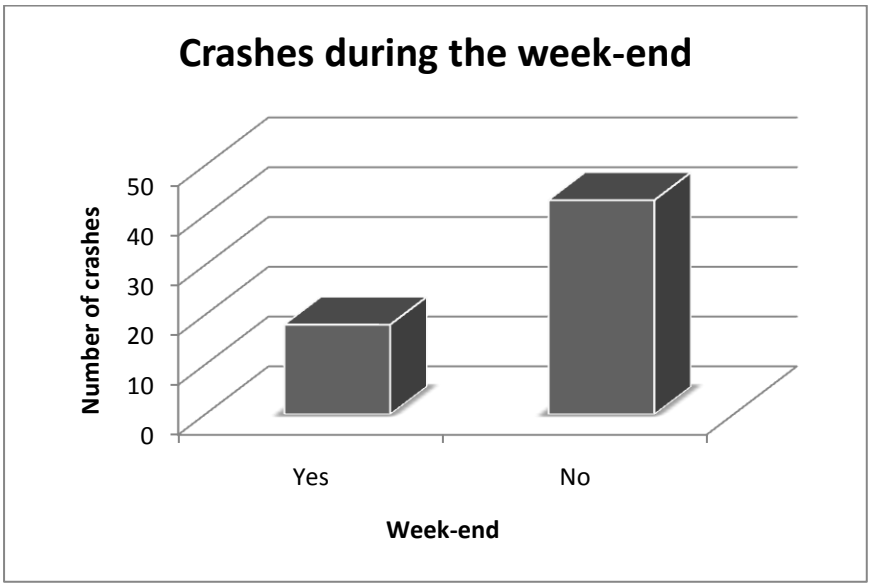

Figure 3.3b: Crashes during the week-end.

Table 3.4a: Seasonal distribution of crashes

Season	Frequency	Percentage (%)
Winter	6	9.84
Spring	8	13.11
Summer	17	27.87
Autumn	30	49.18

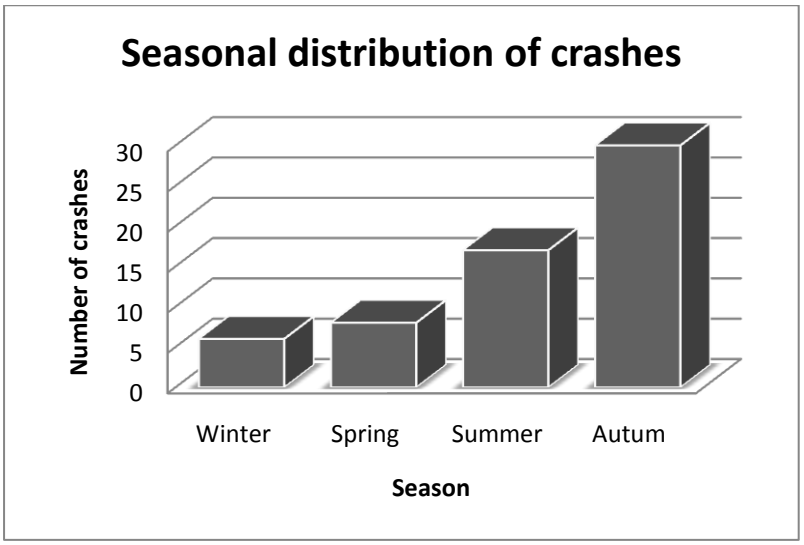

Figure 3.4b: Seasonal distribution of crashes

Table 3.5a: Pedestrian involvement in crashes

Pedestrian	Frequency	Percentage
No	53	86.89
Yes	8	13.11

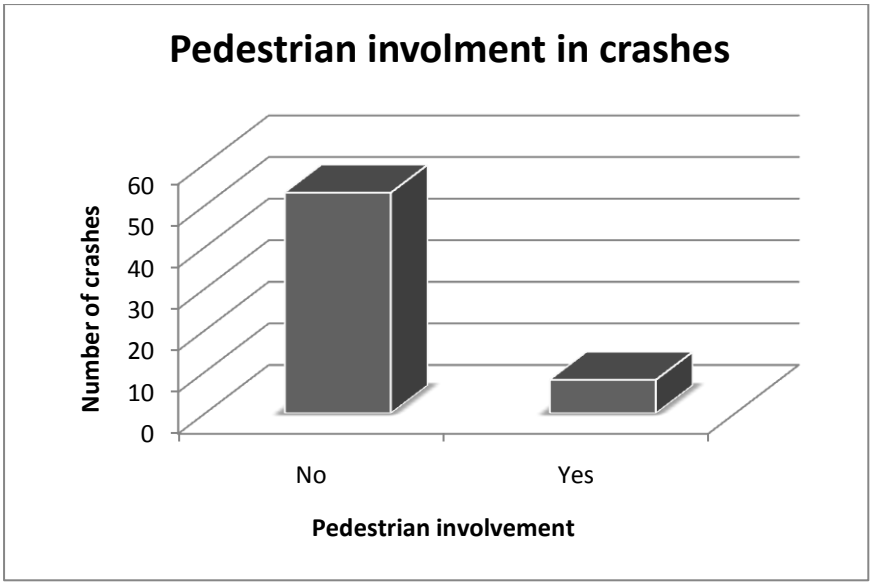

Figure 3.5b: Pedestrian involvement in crashes

Appendix 4: Additional Descriptive Statistics for Intersec_DTLTR

Table 4.1a. Frequency distribution of Collision types

COL_T	Frequency	Percentage (%)
COL_T1	22	11
COL_T2	37	18.5
COL_T3	127	63.5
COL_T4	7	3.5
COL_T5	6	3
COL_T6	1	0.5

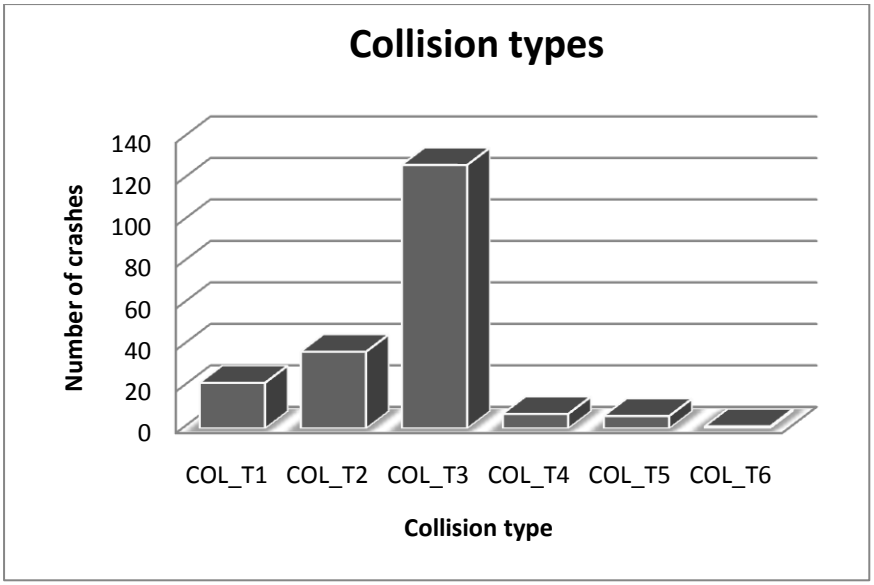

Figure 4.1b: Collision types

Table 4.2a: Frequency distribution of *HOUR*

Hour	Frequency	Percentage (%)
Mornx	40	19.61
L.Mor	35	17.16
After	33	16.18
EvenR	48	23.53
Evenx	25	12.25
Night	23	11.25

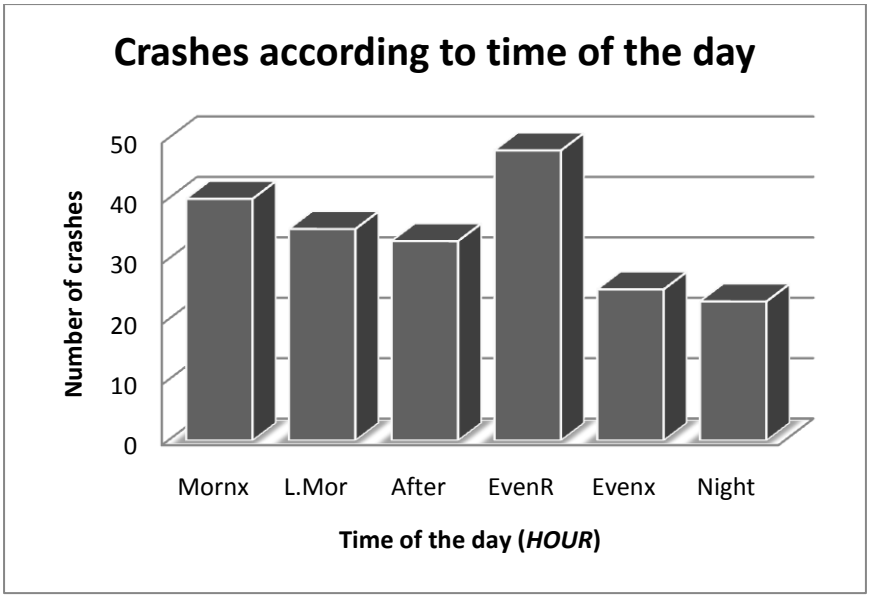

Figure 4.2b: Crashes according to time of the day (*HOUR*)

Table 4.3a: Crashes during the week-end

Week-end	Frequency	Percentage (%)
No	145	71.08
Yes	59	28.92

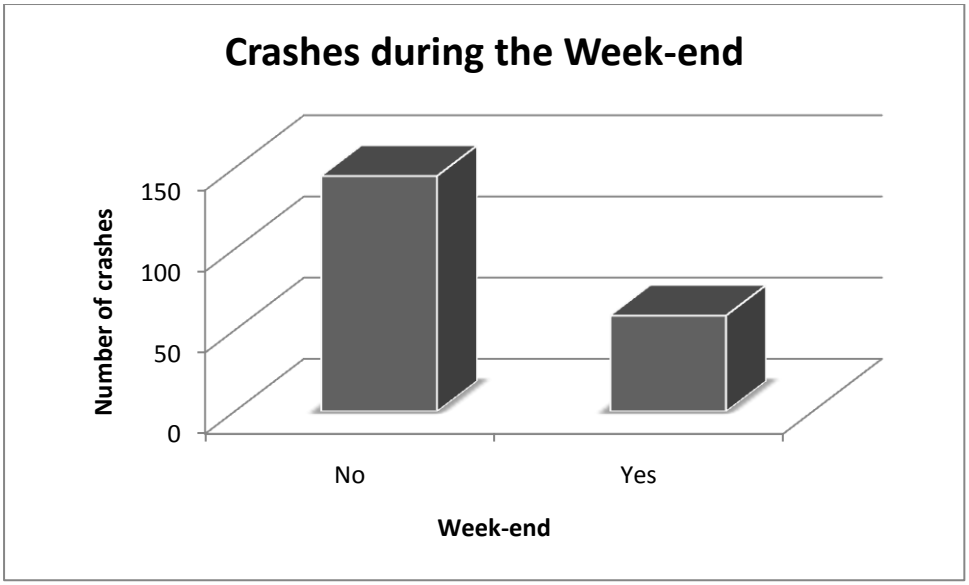

Figure 4.3b: Crashes during the week-end

Table 4.4a: Seasonal distribution of crashes

Season	Frequency	Percentage (%)
Winter	50	24.51
Spring	38	18.63
Summer	55	26.96
Autumn	61	29.9

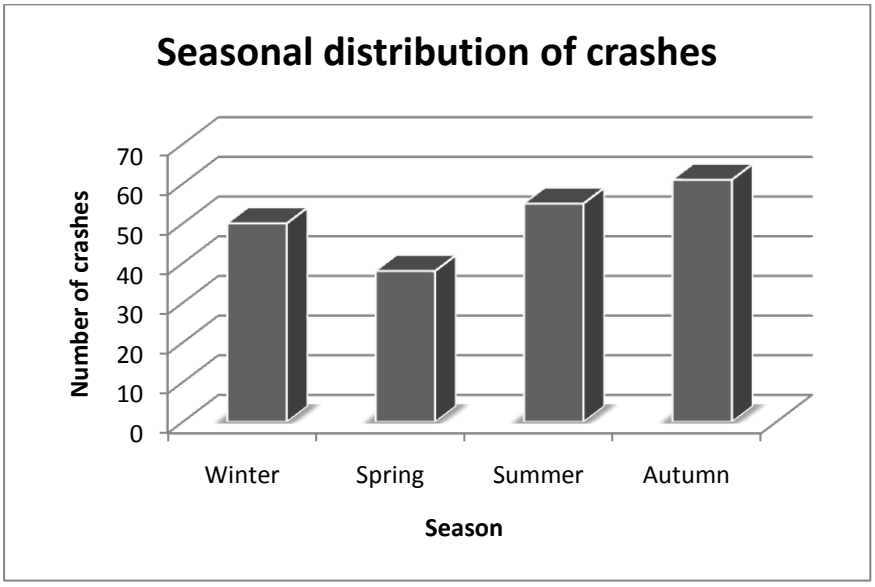

Figure 4.4b: Seasonal distribution of crashes

Table 4.5a: Pedestrian involvement in crashes

Pedestrian	Frequency	Percentage (%)
No	191	93.63
Yes	13	6.37

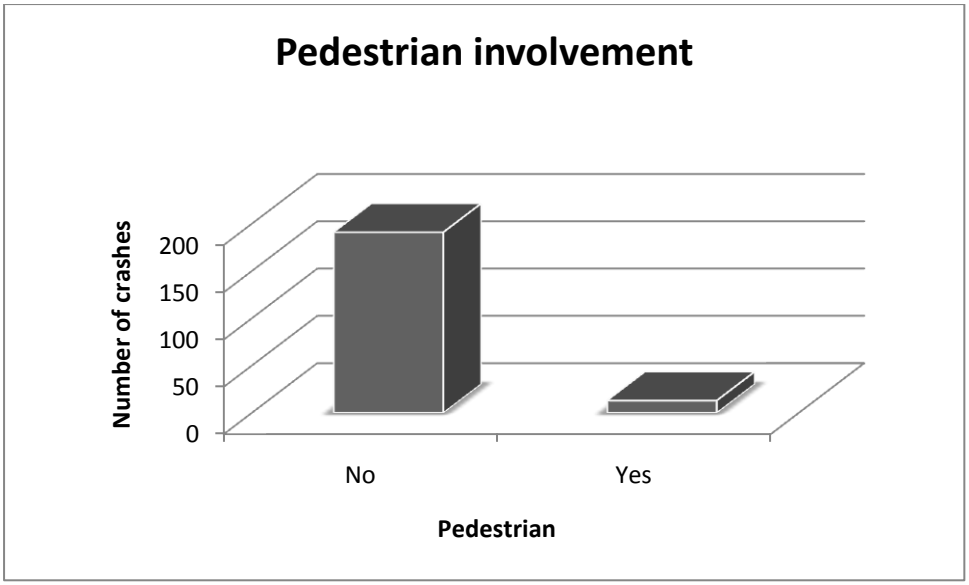

Figure 4.5b: Pedestrian involvement in crashes

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Cluster analysis of crashes on intersection types in Belgium

Richting: **master in de verkeerskunde-verkeersveiligheid**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Ediebah, John

Datum: **21/08/2011**