The passion for science and the passion for music are driven by the same desire: to realize beauty in one's vision of the world.

Heinz Pagels

Acknowledgements

This study was carried out at Hasselt University and has been supported by grant given by the Flemish Government to the Flemish Policy Research Center for Traffic Safety. However, many others have also contributed, all in various ways, to the realization of this thesis.

First of all, I wish to express my sincere gratitude and appreciation to Prof. dr. Geert Wets and Prof. dr. Koen Vanhoof, my promotor and copromotor, for introducing me in the world of data mining and giving me the opportunity to broaden my knowledge on the field of traffic safety. I am also grateful to dr. Tom Brijs, for the time, effort and interest he put in the development of my work. Furthermore, I would like to thank Prof. dr. Lode Vereeck for his experience, encouragement and invaluable support in finalizing this thesis and Prof. dr. Simon Washington and dr. Rune Elvik for their valuable remarks and critical evaluation of my work. Finally, I am also truly grateful to Prof. dr. Dimitris Karlis. His help and inspiring views on statistics have significantly contributed to the results in this thesis.

I would also like to thank all my colleagues at Hasselt University for creating a pleasant and stimulating work environment over the past years. Especially, I am much in debt to my office-mate and true friend Benoît, for his loyal companionship, patience and the many vivid and pleasant discussions, Evy, for her kindred spirit and the thousands of invaluable moments we shared, Elke, for her warmth and support no matter what, and finally Kelly, Lenny and Birgit, for their contagious enthusiasm and true friendship. Thank you all for your support and for never have stopped believing in me and my work.

Also many thanks to my friends and family, in particular my parents, brothers, Tant and Omi, for providing me their support, encouragement and warm home.

Finally, to Wim, my companion in life for over 10 years, thank you for your endless patience and love and to never have made me walk this journey alone.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

This first chapter provides an introduction to the topics covered in this dissertation. More specifically, the issue of traffic safety and road safety engineering as a strategy to address road safety problems is explained. In this context, we introduce the concept of hazardous accident locations and hazardous road location programs. Furthermore, we will present the objectives of this dissertation and provide an overview of the chapters to follow.

1.1 Traffic Safety

In 2002, 47,619 traffic accidents with casualties occurred on the Belgian public roads. In these accidents 1,353 persons were killed, while 8,230 persons were seriously injured and 56,759 persons were lightly injured (National Institute for Statistics, 2002).

Figure 1.1 compares the level of traffic safety in the 25 European member countries for 2002 (Daniëls S., 2005) more specifically, these figures represent the number of road fatalities per million inhabitants. In this list, Belgium is ranked on the 12th position with 128 road fatalities per million inhabitants, which is higher than the European average. Consequently, based on these figures, Belgium has a bad record towards traffic safety in comparison with many other European countries.

Additionally, the steady increase in traffic intensity does not only pose a heavy burden on society in terms of the number of casualties, the insecurity on the roads also has an important effect on the economic costs associated with traffic accidents (Elvik and Vaa, 2004). First of all, there are the non-material costs associated with traffic

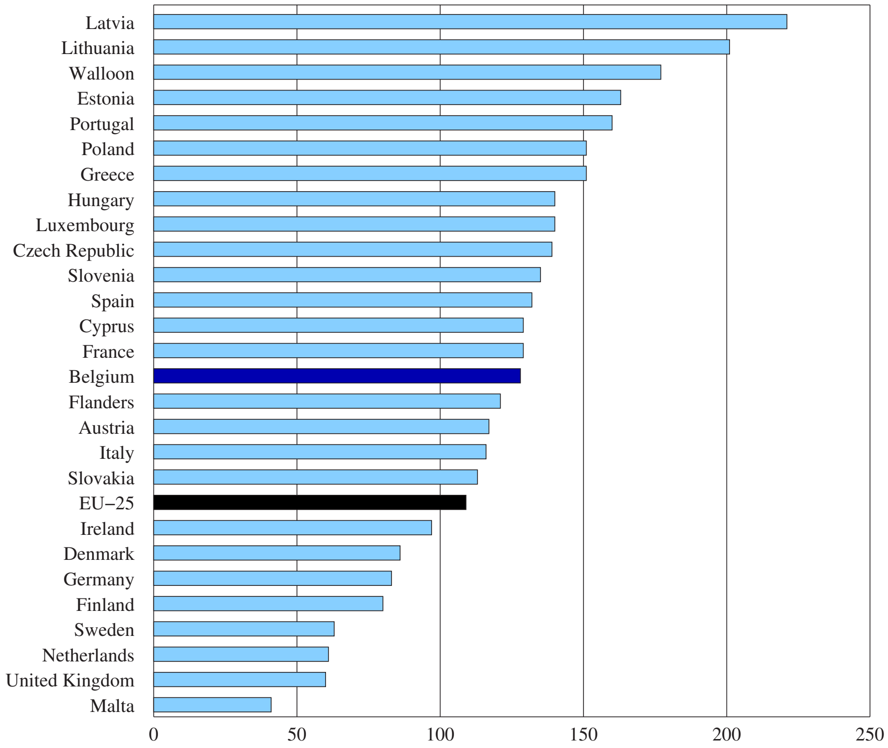

Figure 1.1: Road fatalities per million inhabitants (Source: EU-CARE, NIS)

accidents. These costs concern the pain, grief, reduced joy of life and personal damage influencing the welfare of the victims[1]. Secondly, traffic accidents cause material costs, which can be classified into direct and indirect costs. More specifically, direct costs are related to the accident itself, such as administrative costs, costs of material damage, medical costs and costs as a result of reduced traffic flow. Indirect costs are related to the victims of traffic accidents that are, either temporarily (during recovery) or permanently (in case of death or disability), unable to participate in the production process. Finally, traffic accidents result in costs related to the increased feeling of insecurity on the roads. For Belgium, these total costs that are associated with the insecurity on the roads result in a total macro-economic loss that is estimated for the year 2002, at 12.5 billion Euros (De Brabander and Vereeck, 2005). Accordingly, traffic safety is currently one of the highest priorities of the Belgian and Flemish

[1] The Ministry of the Flemish Community refers to Lindenbergh, Smartengeld, Leiden (1998), Chapter 1

government. More specifically, in the long term the Flemish government has set its goals corresponding to the ‘Zero Vision’, implicating zero deadly or injured road victims on the Flemish territory. However, in the short term, by 2010, the Flemish government aims to have improved its level of traffic safety to such an extent that its arrears compared with the European leaders are cut down to half, taking into account the number of road fatalities per 100,000 inhabitants (Pact of Vilvoorde, 2001). Correspondingly, the Belgian government has set the objective of reducing the number of road fatalities by the year 2010 to 50 percent of the average number of fatalities in 1998, 1999 and 2000. This corresponds with an objective of counting a maximum of 750 road fatalities on the Belgian roads in 2010. When setting this goal, the government has taken into account the increase in flow of traffic that should be expected by this date. In particular, in the period of 1980-1995, the flow of traffic in the European Union increased with 50%. When no measures are taken concerning the means of transport, expectations are that by 2010 traffic will have increased with 33% compared to the traffic level in 1998 (Staten-Generaal van de Verkeersveiligheid, 2001).

1.2 Road Safety Engineering

In order to address the road safety problems, a wide range of possible road safety strategies can be devised. These strategies can be aimed at exposure control, accident prevention, behavior modification, injury control or post-injury management. Road safety engineering falls mainly within the second of these categories, and partly within the third. More specifically, road safety engineering may be defined as a process, based on analysis of road and traffic related accident information, which applies engineering principles in order to identify road design or traffic management improvements that will cost-effectively reduce the cost of road accidents (Ogden, 1996).

The opportunities for road safety engineering in general apply at four levels[2]:

- safety conscious planning of new road networks,
- incorporation of safety features in the design of new roads,
- improvement of safety aspects of existing roads to avoid future problems, and

[2] Ogden refers to Ross Silcock Partnership, (1991), Towards safer roads in developing countries. 220 pp. (Transport and Road Research Laboratory, Crowthorne, UK).

- improvement of known hazardous locations on the road network.

The third and fourth of these applications are closely related but differ in one important respect, namely that one is pro-active, aimed at accident prevention, while the other is reactive, focussing upon remedial treatments at sites which are known hazards, based upon their accident history.

Accordingly, we can state that in general the road safety engineering approach is based on the following elements (European Union Road Federation, 2002):

1. The application of safety principles for accident prevention over new and existing roads.

2. The application of cost-effective measures on existing roads as a basis of accident reduction.

In this context, proactive safety tools such as road safety audits and road safety audit reviews can be implemented (NCHRP Synthesis 336, 2004). These tools have been used by transportation safety professionals since the 1980s. More specifically, a road safety audit is an examination of a transportation project, plan or design roadway, in which an independent, qualified auditor reports on safety issues. The step-by-step procedure of a road safety audit can be performed during any or all stages of a project, including planning, preliminary design, detailed design, traffic control planning, construction, pre-opening, and on existing roads. For an existing roadway section or intersection, either just before opening or already open to traffic, the road safety audit is effectively a review and is discussed as a road safety audit review.

However, traditional safety planning practice often includes reactive rather than proactive elements. A problem is identified, primarily through analysis of accident data, and an appropriate enforcement, education or engineering countermeasure is implemented. Indeed, every year, states, counties, regions and municipalities spend considerable amounts of resources on trying to reduce crashes by reconstructing and improving the roads. In contrast with road safety audits, which are in essence aimed at crash prevention, this approach is aimed at crash reduction.

Authorities should opt for reactive or preventive measures depending on the situation they have to respond to in terms of road safety.

1.3 Hazardous Accident Locations

1.3.1 Different Aggregation Levels

As explained in the previous section, road safety engineering can be applied at different levels. Indeed, accident histories usually need to be aggregated in order to have some confidence in the beneficial effects of remedial measures. For example, a single accident at a site is a poor indicator of what may happen in the future, but if there are several accidents of the same type occurring at the site, one can be more confident that a remedial measure, focussed on that particular accident type will be effective. Examples of appropriate aggregations include (Ogden, 1996):

- accidents clustered at intersections or on short lengths of a road
- accidents clustered along routes or sections of routes
- accidents clustered within an area
- groups of accidents for which there are known effective treatments, occurring across several sites
- a series of accidents with common features, such as road features (e.g. bridges), vehicle features (e.g. bicycles), road user features (e.g. pedestrians) or contributory features (e.g. driver fatigue)
- a series of 'high profile' accidents such as those involving vehicles carrying dangerous goods, or accidents at railway crossings

Consequently, four basic approaches for reducing crashes by applying engineering treatments or countermeasure can be discerned (The Bureau of Transport and Regional Economics of Australia (2001)):

- Single sites or black spots: treating specific sites or short sections of road
- Route action: applying known remedies on a route with an abnormally high crash rate
- Area-wide action: applying several treatments over a wide area
- Mass action : applying a known remedy to locations with common crash problems or causal factors

In this research, we will focus on the first of these approaches, more specifically, treating dangerous sites or short sections of roads.

1.3.2 Correction of Hazardous Locations

Literature points out that there is no universally accepted definition of what should be considered as a dangerous location (Hauer, 1996). In general, locations are classified as hazardous sites after an assessment of the level of risk and the likelihood of a crash occurring at a location. Locations that have an abnormally high number of crashes are then described as crash concentrated, high hazard, hazardous or black spot sites (Bureau of Transport and Regional Economics of Australia, 2001). In order to improve safety at these sites, a hazardous road location program, also called black spot safety work, can be defined. This reactive process aims to identify locations within the road system which have an unacceptably high incidence of road accidents, in order to develop appropriate treatments to reduce the costs of these accidents (Ogden, 1996).

More specifically, the overall goal of a hazardous road location program is to identify locations at which there is not only an inherently high risk of accident losses but also an economically justifiable opportunity for reducing this risk. Additionally, it aims at identifying countermeasure options and priorities which maximize the economic benefits from the hazardous road location program[3]. For simplicity, a hazardous road location program can be summarized in its most simple form as shown in figure 1.2 (Ogden, 1996).

Analogously with this figure, and based on Schlüter (1997) and Vistisen (2002) typical procedures for hazardous site correction can be divided into three basic tasks:

1. The identification and ranking of hazardous locations. This results in a list of sites with promise ('Identification Phase').
2. Prioritizing these sites by diagnosing the problems at identified locations and determining potential remedial treatments in order to identify cost-effective safety improvement projects('Investigation Phase').
3. The appraisal of alternative treatments followed by implementation of the best treatment if sufficiently cost-effective. To evaluate the effect of treatment, before and after studies need to be conducted. ('Program Implementation Phase')

In this dissertation, we will focus on the first and second element of these steps, namely the identification and investigation of hazardous accident locations.

[3] Ogden refers to Sanderson and Cameron (1986). Identification of hazardous road locations. Proc 13th Australian Road Research Board Conference 13(9), pp. 133-147

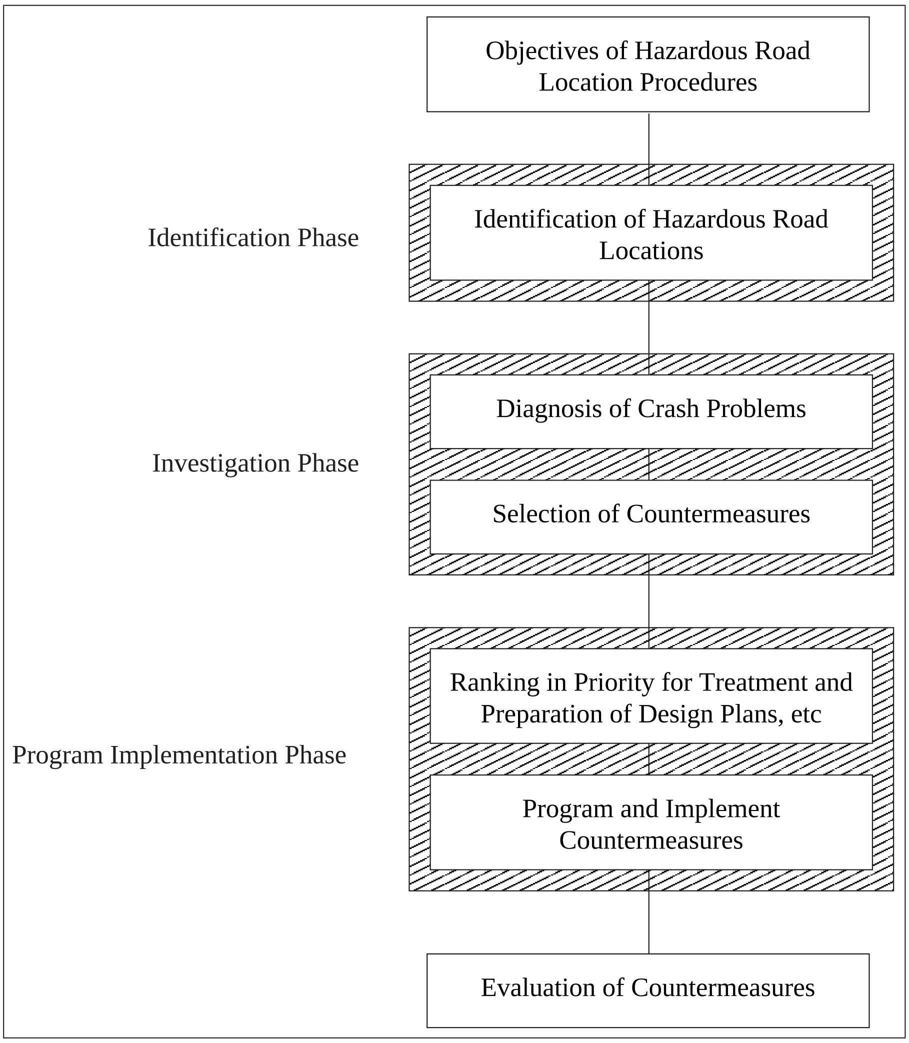
Objectives of Hazardous Road Location Procedures
Identification Phase
Identification of Hazardous Road Locations
Investigation Phase
Diagnosis of Crash Problems
Selection of Countermeasures
Program Implementation Phase
Ranking in Priority for Treatment and Preparation of Design Plans, etc
Program and Implement Countermeasures
Evaluation of Countermeasures

Figure 1.2: Hazardous road location program elements (Source: Ogden, 1996)

1.4 Research Objectives

1.4.1 Identification and Ranking of Hazardous Sites

For the identification and ranking of hazardous sites, different criteria can be used. In general, based on Ogden (1996) and Taylor and Thompson (European Union Road Federation, 2002), we can discern seven methods that can be used to identify dangerous sites on the road network, each with different order of importance and precision:

1. Accidents Frequency: in this method only the number of accidents (or accidents per unit length of road) in a given period are considered using a predetermined threshold value. The disadvantage of this first method is that traffic volume and accident severity are not taken into account, nor the randomness of accidents.

2. Hazard Potential Ratio: here, the number of accidents and the traffic volume are considered together. These ratios are usually expressed in terms of accidents per million vehicle kilometers. Then, road sections are considered as dangerous using benchmark criteria.

3. Joint Method with Accident Frequency and Accident Risk Ratio: this method aims at developing a selection of road sections according to their levels of accident frequency. Next, it aims at prioritizing the measures to be taken depending on the road section's accident risk ratio.

4. Confidence Interval Method: this method, sometimes referred as the rate quality control method, is based on the application of a statistical test with the aim to determine if the particular accident risk ratio of a road section is significantly higher than the mean value of this ratio. This mean is previously determined for locations with similar characteristics.

5. Method of the Accident Severity Ratio: road sections are classified according to their gravity rate which is calculated using weighted coefficients for different types of casualties.

6. Risk Rate Method: this method defines for each of the different characteristics influencing the accident risk ratio of road sections a risk function. The value of this risk function expresses the accident risk connected to the road section depending on its characteristics. Consequently, road sections that have a risk function value higher than a predetermined value are considered as dangerous.

7. Inventory of the Accident Risk Elements in the Road: the purpose of this method is to identify road sections that present some high potential accident risk characteristics without having a particular high accident rate. Before identifying these road sections, a definition of the characteristics considered as dangerous is developed. The danger represented by these characteristics is measured through the expected accident risk ratio.

These definitions indicate that there is a wide range of methodologies available, ranging from simple models based on actual counts to advanced statistical models based on estimates to identify dangerous locations.

In this dissertation, we will investigate how hazardous accident locations are currently identified and ranked in Flanders (the Flemish speaking community of Belgium). At the moment, in Flanders, 1,014 accident locations are considered as dangerous (Ministry of the Flemish Community, 2001). These accident sites are selected by taking into consideration the number of accidents at the location and weighing the number of light, serious and fatal injuries at each location. More specifically, for each location where in the last three years, three or more accidents occurred, a priority score is calculated by summing up the number of injuries at these locations while using weighing values of respectively 1, 3 and 5 for the injury types light, serious and fatal. The accident sites with a priority score of 15 or more, are then considered as dangerous. To improve traffic safety on these locations, the Flemish government will each year, starting in 2003 for a period of five years, invest 100 million EURO to redesign the infrastructure of the 800 accident locations with the highest score. In this dissertation, a sensitivity analysis is performed to investigate the strengths and weaknesses of this approach to rank and select dangerous accident sites. More specifically, we will investigate how big the impact would be on the current selection of dangerous accident locations when different weight values are used. Furthermore, the effect of giving weight to the severity of the accident instead of to all the injured occupants of the vehicle is investigated. Finally, effects of using the expected number of accidents (using Bayesian estimation techniques) instead of using historic count data are evaluated. By investigating whether these different ranking criteria have a big impact on the selection of dangerous accident locations, we want to sensitize government to carefully choose the criteria for ranking and selecting dangerous accident locations.

Indeed, the ordered list of dangerous locations is important as locations are generally selected by working down the list until the allocated resources are exhausted for the detailed examination (i.e. the diagnosis and identification of potential treatments), and perhaps, subsequent treatment of locations. Different list orderings may lead to a different set of locations being examined in detail. An inappropriate ordering of locations, therefore, could lead to a truly dangerous location not being examined and considered for treatment, e.g. false negatives (locations not identified as hazardous because the number of accidents happened to be abnormally low) while money is invested in other truly safe sites, unjustly identified as dangerous, e.g. false positives (locations that have a high recorded number of accidents mainly as a results of chance) (Schlüter et al., 1997).

1.4.2 Investigation of Hazardous Sites

As explained in figure 1.2, after the hazardous sites are identified, it is necessary to carefully examine the nature of the safety problem at the sites with a view to identifying whether and how those problems can be dealt with through road or traffic remedial measures. This allows to set a diagnosis of the problems at the identified locations and to determine potential remedial treatments (Ogden, 1996).

Indeed, hazardous sites are difficult to analyze, because there are many factors associated with them (Bureau of Transport and Regional Economics of Australia, 2001). In order to develop effective countermeasures to reduce the number of accidents at dangerous locations, one should therefore properly and systematically relate accident frequency and severity to a large number of variables such as roadway geometries, traffic control devices, roadside features, roadway conditions, driver behavior or vehicle type (Kononov and Janson, 2002). Furthermore, analyzing road accidents allows the identification of crash patterns that may arise at sites that may not reveal elevated crash frequencies or rates. The patterns could emerge as crash outcomes (e.g. over-represented rear-end collisions) or simply elements in common (e.g. night-time crashes)(see e.g. Shankar et al. (1995) and Kim et al. (2006)).

For this purpose, since a few decades, traffic accident data are registered and analyzed to support the traffic safety policy in Flanders. Additionally, as mentioned in the Design Report of the Status Questionis Flanders (2002), these accident data are located in a Geographical Information System (GIS). When necessary, this pinpointing is done with the help of official police reports. The GIS tools allow for the local authorities to analyze the located accident data in Access databases. However, this research is usually limited to the analysis of the evolution of accident data in relation to certain infrastructure parameters such as the number of accidents in the built-up area or the number of accidents on numbered roads. The importance of the environmental factors in relation to other relevant factors needs to be further researched to better identify which combinations generate higher concentrations of accidents.

The following risk factors have been elaborately adopted in the literature for explaining accident involvement and accident severity: course of the accident (e.g. vehicle manoeuvre, driver action), traffic conditions (e.g. traffic volume, dynamics, speed regulation), environmental conditions (e.g. light condition, road surface condition, road geometry), human conditions (e.g. driver age, occupant age, driver sex, driver condition (e.g. alcohol, fatigue, illness, seating position, seat belt use) and vehicle conditions (e.g. vehicle mass, vehicle size) (Nassar, 1996).

Furthermore, several attempts are found for explaining the spatial variation of road unsafetyness at several levels of spatial aggregation (see Flahaut, 2004a, 2004b for a review). In this dissertation, we will follow an exploratory approach. More specifically, we are interested in finding out which factors are associated with the hazardous accident locations. Therefore, we will profile the accident locations in terms of accident related data and location characteristics to provide new insights into the complexity and criteria that play a significant role in the occurrence of traffic accidents.

For this purpose, in the past, statistical models have been widely used to analyze road crashes in order to explain the relationship between crash involvement and traffic, geometric and environmental factors (Lee et al., 2002). However, Chen and Jovanis (2002) demonstrate certain problems that may arise when using classic statistical analysis on data sets with large dimensions such as an exponential increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables. This is where data mining comes into play. Data mining can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from large amounts of data (Frawley et al., 1991). The use of data mining methods can therefore be particularly useful in the context of large data sets on road accidents.

Therefore, in this dissertation we will investigate how data mining and statistical techniques can be implemented to identify and profile dangerous accident locations in terms of accident related data and location characteristics. Furthermore, these techniques will be used to develop models, which will provide new insights into the criteria that aim to explain to probability of the occurrence of a traffic accident.

1.5 Dissertation Outline

Chapter 2 of this dissertation introduces the main problems encountered with the identification and ranking of hazardous accident locations such as accident migration and regression to the mean. Additionally, the issue of defining the best length of a dangerous road segment and the problem of underreporting of road accidents are explained. Next, an overview is presented of several statistical models that are described in literature to model road accidents taking into account these initial considerations. Then, several alternative methods that are used to identify and rank hazardous accident locations are described. Furthermore, the use of before and after studies to estimate the effect of treatment on these sites is briefly discussed in this text. Finally, an overview of the hazardous sites approach in Flanders is presented.

Next, chapters 3, 4 and 5 will contain the main research contributions of this dissertation. More specifically, in chapter 3, a sensitivity analysis is performed on the currently used method to rank and select dangerous accident locations in Flanders. In particular, we will evaluate how big the impact would be on the current selection of hazardous accident locations when respectively different weight values are used, when weight is given to the accident instead of to all the injured occupants of the vehicles, and when the expected number of accidents, estimated from a hierarchical Bayes model, instead of the historic count data are used to rank and select accident sites. Additionally, an optimization model is developed in order to automatically generate the 'optimal' weighing values for the ranking and selection problem. In chapter 4, dangerous accident locations are profiled in terms of accident related data and location characteristics using an association algorithm and model based clustering. More specifically, we will use these techniques to profile respectively different accident types, high frequency accident locations, black zones and high risk road clusters. Then, in chapter 5 we will investigate whether the dangerous accident sites in Flanders tend to migrate over time using accident data from different time periods.

Finally, chapter 6 is reserved for conclusions and an overview of topics for future research.

Chapter 2

Hazardous Sites Analysis Methods: Background

There is a fairly extensive literature focussed on methods for the identification of ‘black spots’, ‘sites with promise’, ‘high risk’ or ‘hot spots’ covering an exhaustive set of issues (Cheng and Washington, 2005). Some papers address regression to the mean issues, while others address crash outcome versus total crash modeling. Some discuss the application of Bayesian methods, while others try to make sense of cross-sectional data. This chapter starts with the discussion of some initial considerations on the main problems that are encountered with when analyzing hazardous accident locations. In this context, several statistical models are described in literature to model the accident frequency and accident severity. Additionally, a literature review on the most important methods that are used by different authors to identify and rank hazardous locations is presented. Furthermore, the use of before- and after studies to estimate the effect of treatment on these sites is discussed. Finally, an overview of the hazardous sites approach in Flanders is presented[1].

[1] Part of this chapter has been published as follows:
Geurts, K. and Wets, G. (2003). Black Spot Analysis Methods: Literature Review. Rapport Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit, RA-2003-07, Diepenbeek

2.1 Initial Considerations

Apart from random issues in road accidents, the main problems encountered with the identification of hazardous accident locations are accident migration and regression to the mean (European Road Federation, 2002). In this context, several statistical models are described in literature to model the variation in accident frequencies. Furthermore, no clear indication exists of what the best length of a dangerous road segment should be. Finally, the issue of underreporting of road accidents will play a role in the identification of hazardous accident sites. These issues are briefly discussed in the following paragraphs[2].

2.1.1 Random Issues in Road Accidents

The annual number of accidents occurring on a road section varies each year. This is explained by the inherent accident risk of the road section. The annual random accident rate is linked to the accident rate's own nature. It also depends on a range of factors that cannot be forecasted. Consequently, if the random nature of the accident rate is not considered, measures and funds allocated to safety issues will probably be misused. Therefore, it is important to analyze the road network during:

- A representative time period in which a statistically sound number of accidents occur. This means that the analysis period must be long enough to yield representative accident samples. Following that principle, a large number of studies have been made and it has been generally agreed that in most cases a period of three to five years is sufficient to guarantee the reliability of the analysis (see for example Cheng and Washington, 2005).
- A time period in which no substantial modifications of the infrastructure or the environment have been made.

2.1.2 Regression to the Mean

A consequence of the random character of a road section's accident rate is the regression to the mean. If black spot sites are chosen for treatment solely on the basis of their high crash record in a certain year, the chosen sites may genuinely be very hazardous. However, it is also possible that the high crash rates observed at some

[2] This section is mainly based on the dissertation of Vistisen (2002)

sites may be due to chance, or a combination of both chance and a moderately hazardous nature. In other words, a traffic accident black spot may be due to adverse road conditions that require extreme care. But it may also be nothing more than a meaningless cluster of accidents which occurs randomly and has no intrinsic meaning. These sites are likely to have fewer crashes in a subsequent period even if no treatment is carried out, because the number of crashes will tend to gravitate towards the long-term mean value. Under these conditions, the effect of any treatment is likely to be over-estimated. This effect is most commonly called the regression to the mean effect and is sometimes known as the bias by selection effect (see Davis (1986) for a general description).

According to Hauer (1997), we can say that for an accident location, the observed accident count K always fluctuates around some unknown expected value κ. If K was a good estimator of κ, entities which recorded K accidents in one period would then record, on average, K accidents in a subsequent period of equal duration (if κ remained unchanged). However, empirical evidence has shown that, even if K in one period happens to be larger than κ, the best guess about the magnitude of K in the next period is still κ. Consequently, if K is not a good estimator of κ, a better method of estimation needs to be suggested. The empirical Bayes approach, introduced in the next section and discussed into more detail in the following chapter, will fulfil this need (Hauer, 1997).

2.1.3 Modeling Variation in Accident Frequencies

To account for the random character of accident counts, in literature a number of statistical models have been used to estimate accident rates and/or accident frequencies at a specific location over a given interval of time. According to Nassar (1996), the following models can be discerned.

Foldvary (1979) and Jovanis and Delleur (1983) have used simple models using mean and variance. These models are used to study variations in accident rates for different levels of exposure. These models, however, are not able to incorporate the effect of risk factors on accident involvement.

In Oppe (1979) and Ceder and Livneh (1982) multiple linear regression models are used. In these models the dependent variable (either number of accidents or accident rate) is a function of a series of independent variables such as speed or traffic volume. Accident occurrence in these models is assumed to be normally distributed. However, these models generally lack the distributional property that is necessary to describe adequately the random and discrete vehicle vehicle accident events on the road and they are inappropriate for making probabilistic statements about accident occurrence. Therefore, other types of models were proposed for estimating traffic accident events.

To account for the probabilistic nature of accident occurrence, Saccomanno and Buyco (1988) and Blower et al. (1993) have used a Poisson loglinear model to explain variations in accident rates. The occurrence of accident events is rare and random in nature, which fits the Poisson distribution well theoretically (Land et al., 1996). More specifically, this Poisson regression model is especially suitable for handling data with large number of zero counts. However, this model could be inappropriate in certain situations where the value of the variation deviates from the value of the mean in the observed accident counts. When the variance is larger than the mean, this is termed overdispersion (McCullagh and Nelder, 1989).

This discrepancy can be overcome by an additional assumption on the expected mean, which is a gamma distribution assumption. In brief, it is assumed that the number of accidents follows a Poisson distribution with rate parameter lambda. The rate parameters are assumed distributed across the population of locations according to a gamma distribution. The distribution of the accident counts at the population level is then given by the negative binomial distribution (Cox, 1983; McCullagh and Nelder, 1989; Land et al., 1996). The negative binomial regression model is particularly useful in accounting for the overdispersion (Land et al., 1996). Several authors such as Maycock and Hall (1984), Hauer and Persaud (1987), Persaud (1990), Miaou (1994), Shankar et al. (1995), Maher and Summersgill (1996), Kulmala (1995), Hauer (1997), Tunaru (1999), Abdel-Aty and Radwan (2000) have used negative binomial regression models.

Persaud (1990) incorporated an empirical Bayesian adjustment in his negative binomial model arguing that adjusting historical data by statistical estimates yields improved predictability. Ever since, many applications used in some ways an Empirical Bayes approach (Schlüter et al., 1997; Van den Bossche et al., 2002). Recently, Cheng and Washington (2005) illustrated by means of an experimental evaluation of hot spot identification methods that the Empirical Bayes technique significantly outperforms ranking and confidence interval techniques. Another form of the Bayesian approach, called the hierarchical Bayesian model, has also been proposed in literature. These methods can handle the uncertainty and the great variability of accident data and produce a probabilistic ranking of the accident locations. However, the use of these models in traffic safety is less widespread (Schlüter et al., 1997; Van den

Bossche et al., 2002).

The issue of extra-Poisson variation in road accident data was addressed further by Hauer and Persaud (1987). They extended the generalized linear model of Maycock and Hall (1984) who allowed for systematic differences in accident frequencies between characteristics such as traffic flow and various geometric variables. In their model, the Poisson mean was allowed to vary between locations beyond what may be explained by differences in characteristics. The Poisson-gamma generalized model set up in Hauer and Persaud (1987) has been supported by Maher and Summersgill (1996), and is now widely accepted (see e.g. Kulmala, 1995; Hauer 1997). A few extensions to this model do have been pursued. As an example, in Tunaru (1999), the parameter in the gamma distribution as well as the regression variables are modeled as random variables themselves, thus adding several levels to the model. However, most modeling of traffic accidents use a generalized linear model with negative binomial error structure, and no specific modeling of dispersion effects (see e.g. Abdel-Aty and Rawan, 2000).

In Lord et al (2004), it is explained that Zero-inflated models have also been used in traffic safety studies for modeling crashes for various applications. Given the characteristics of these models, most of these authors have automatically assumed that crashes must follow a dual-state process, which implies that entities (e.g., intersections, road segments, pedestrian crossings, etc.) exist in one of two states: perfectly safe and unsafe. Although zero-inflated models do offer improved statistical fit to crash data in many cases, the authors argue in that the inherent assumption of a dual state process underlying the development of these models is inconsistent with crash data. However, as was shown with both empirical and simulated data, excess observed zeroes are caused by the following four issues: (1) spatial or time scales that are too small; (2) under or misreporting of crashes; (3) sites characterized by low exposure and high risk; and (4) important omitted variables describing the crash process. Accordingly, at risk when using zero-inflated models is the possibility that misinterpretations about safe and unsafe intersections/road segments/etc. will result. Someone might falsely believe that certain engineering investments as predicted by the zero-inflated models will lead to inherently safe locations. Therefore, theoretically defensible solutions for modeling crash data with excess zeroes include changing the spatial or time scale of analysis, including unobserved heterogeneity terms in Negative Binomial and Poisson models, improving the set of explanatory variables, and applying small area statistical methods.

In the models described above, only the reported number of accidents in the observation period is used and location characteristics are thus modeled as constant within this period. In practice, these characteristics (in particular traffic flows) often change over time. In order to account for such changes, one may wish to divide the observation period into sub-periods. However, because accident counts, in different sub-periods at the same location, depend on the same site-specific conditions not reflected in the location characteristics, they are not independent (see Maher and Summersgill, 1996). This poses difficulties in estimating the models, as accident counts no longer are independently negative binomially distributed. For example, in Persaud (1994) each year accident's count is used as separated records. However, Persaud made the unrealistic assumption of independence between yearly accident counts at the same time. In Vistisen (2002), disaggregated accident models on sub-periods of one year are used with the assumption that yearly accident counts at the same location are dependent. The random variation in accident frequencies is described by a hierarchical Poisson-gamma distribution, but the Poisson mean is separated into a fixed and a dispersion part (a parametrization also used in Hauer, 2001).

Finally, Jovanis and Chang (1989) have used survival theory models. These models predict the probability of a vehicle being involved in an accident at time t given that the vehicle had survived until that time. Since the use of these models requires specifically collected data, this approach has not been widely adopted by other researchers in this field. Indeed, what distinguishes survival data from other kinds of data, is the inevitable presence of incomplete observations (Petersen et al., 1996). Often, practical restrictions prevent the observation of the terminal event of interest for every individual in the sample, in which case the available piece of information is a right-censoring time, a period elapsed in which the event of interest has not occurred (in this case, no accident occurred). Additionally, in traffic safety analysis left censoring occurs because we don't observe the true starting point for the vehicles. For example, if the start is when they last crashed, we probably observed them after that fact. Right censoring occurs because those that did not crash in our study may eventually crash after our study, and so we don't observe their final event.

In this dissertation, we followed the approach of Brijs et al. (2003), who proposed a multivariate hierarchical Bayes approach for ranking accidents sites taking into account the number of accidents, the number of fatalities, and the number of light and severely injured casualties for a given time period for each site. This is done by using a 3-variate Poisson distribution that allows for covariance between the number of lightly, seriously and fatally injured casualties. This will be explained in the third chapter of this dissertation.

2.1.4 Simulated versus Empirical Data

Besides the development of statistical models to account for the random character of accident counts, the use of simulated data instead of empirical data is also proposed in literature. More specifically, Cheng and Washington (2005) use experimentally derived simulated data to evaluate different hot spot identification methods. The authors argue that when analyzing real data, i.e. crash counts, the analyst never knows a priori which sites are truly hazardous and which sites happen to have experienced random up fluctuation in crashes during a period of observation. This leads to the very difficult situation of trying to count false positives and negatives without knowing which sites are truly safe and 'unsafe'. In contrast, in a simulation it is possible to establish a priori sites that are hazardous and assess whether hot spot identification methods can correctly identify them. The simulation approach, however, requires considerable care in 'constructing' the crash data so that they are convincingly similar to empirical crash data. In particular, properties of the observed crash data are used to generate simulated crash frequency distributions at hypothetical sites. Additionally, the authors note that simulated data suffer from lack of realism encountered in 'uncontrolled' observational settings.

2.1.5 Accident Migration

During the development of black spot measures, a relation has been observed between a black spot disappearance and the development of a new black spot near the previous one. Indeed, black spots tend to migrate over time to another location. This phenomenon of accident migration occurs when a black spot is not managed properly: on the one hand the accident number in the existing black spot decreases, while, on the other hand, usually in a road section nearby, the accident number suddenly increases. In fact, accident migration represents a decrease in the black spot management efficiency (European Road Federation, 2002).

This can be explained by the fact that even localized treatments may have non-localized consequences (Hauer, 1997). One such common consequence of treatment is traffic diversion. Clearly if treatment on one location causes changes in traffic elsewhere, safety elsewhere will change too. However, accident migration can also take more subtle forms. For example, in a neighborhood where almost all intersections have been converted from two-way to all-way stop control, people tend to make the mistake of expecting all vehicles to stop, also at the few intersections which remain under two-way stop control. This is a plausible mechanism for accidents migrating from intersections covered to all-way stop control to those which remain unconverted.

Consequently, to prevent accident migration, it is important to determine the target accidents for a treatment (Hauer, 1997). These target accidents are all those that can be affected by the treatment. It follows that a convincing evaluation of the safety effect of a treatment requires a good understanding of the process by which the accidents are generated and avoided.

2.1.6 Sites, Routes and Areas

As explained in the previous chapter, hazardous site analysis involves an examination of the accident patterns at a specific location, such as an intersection, a short length of road or a specific road feature. For the purposes of analysis, it is therefore necessary to define the road length, or in the case of intersections, to be specific about the definition (Ogden, 1996).

In practice, methods developed for identifying accidents concentrations often apply to hot spots (also called black spots, hazardous locations, sites with promise etc.) which are pinpoint concentrations of road accidents that often migrate over time (see e.g. Silcock and Smyth, 1985; Maher, 1990; Nguyen, 1991; Joly et al., 1992; Hauer, 1996; Thomas, 1996 or Vandersmissen et al., 1996). However, although the term 'spot' suggests a precise location, it also often used to refer to sections of roads, but in most studies its length is not justified and not controlled.

Furthermore, in literature, the awareness of spatial interaction between contiguous accident locations arises. However, since in many countries, accident data are available on a single accident basis, this makes all types of spatial aggregation possible: aggregation on a point, on a line, or on a surface. The most relevant aggregation unit for accident data is the road segment. However, no clear indication exists of what the best length of a dangerous road segment should be, nor or whether an optimal length can be defined. (Flahaut et al., 2003). For example, in some empirical accident studies, the road section is divided into road segments either of constant length (e.g. 0.1 mile, 1 mile, 1 km) or of variable length. The size of the segment length is often not discussed, and sometimes even not clearly defined (Thomas, 1996).

Since in Belgium, the location of accidents is only accurately known for the 'numbered' roads, i.e. highways, national and provincial roads linking towns because there is a stone marker every hectometer, the identification of dangerous accident locations is related to roadway segments of numbered roads with a length of 100 meters. Furthermore, each intersection is considered as a possible dangerous site. Accidents occurring in the direct neighborhood of an intersection (within 50 meters) are also incorporated in the calculations of this intersection. This means that the accident locations that are considered as hazardous sites are either roadway segments of 100 meters or intersections.

2.1.7 Underreporting of Road Accidents

One of the most important gaps in the existing accident data provision, and thus inevitably affecting the identification of hazardous accident locations, is in the area of underreporting of road accidents (Hauer and Hakkert, 1988). Indeed, a substantial number of road accidents is not reported to the police, especially in the case of accidents with only material damage or light injuries. The level of this underreporting varies considerably from region to region, but is significant in almost all countries (Asian Development Bank, 2003).

Consequently, in several EU countries, mostly in the northern and western regions, clinical hospital data on traffic injuries are linked with the police reported accident data on a national or regional level (Thomas et al, 2003). This serves two purposes: (1) establishing the underreporting of registration of injury accidents by the police and (2) adding the detailed injury information to the registered data of accidents. In this context, Thomas et all (2003) state that it is generally believed that almost all fatalities are registered, but a German study[3] on linked hospital and police data estimated that up to 5 % could be missing from police data. A French study[4] for the region of Lyon established that as many as 12 % of fatalities were underreported in the official police based registration. In Nordic countries, Great Britain, Germany and the Netherlands (Ministerie van Verkeer en Waterstaat, 2001), several studies on linked hospital and police data[5] have revealed that many injuries from single vehicle

[3] The authors refer to Metzner. G. (1992). Retrospectivanalyse todlicher Verkehrsunfalle unter dem Aspekt einer Fehlerfassung der Verkehrstoten. Zeitschrift fur Verkehrssicherheit: pp. 150-151. Koln.

[4] The authors refer to ETSC (2001). Transport Accident and Incident Investigation in the European Union. European Transport Safety Council, Brussels. Laumon, B., Martin J.L., Coller P., Verney M. P., Chiron M., Ndiaye A., and Vergnes I. (1997)

[5] The authors refer to OECD (1994). Under-reporting of road traffic accidents recorded by the

accidents and injuries of pedestrians and cyclists are underreported to a varying extent in the official road accident registration systems of these countries. On average between 20-40 % of all serious injuries are not reported, while the largest underreporting with respect to all (light and serious) injuries is generally observed for cyclists. Up to 80 % of injured cyclists in traffic accidents are not reported.

A case study on the treatment of road accident casualties by emergency services and general practitioners in the region of Antwerp (Belgium)[6] has shown that no less than 55% of the total number of traffic casualties is not reported in the official figures of the National Institute for Statistics (Lammar, 2003). Moreover, one should take into account that this number does not even include the number of casualties that chose not to receive any treatment for their injuries, nor the number of patients that did not report their injuries to be due to a road accident).

From these figures, it is evident that the statistical analyses and the monitoring of developments from the injury accident data in the national databases will be misleading, unless detailed corrections for fairly well known underreporting percentages are made. Police reports would underestimate the magnitude of injuries and distort any evaluations of preventive initiatives (Nakahara and Wakai, 2001). Therefore, it is urgently recommended that national studies on the underreporting of injuries are periodically performed in every country of the EU. The aim is that correction factors can then be applied to the types of injury data to obtain reliable information on road injuries. This would allow the correct estimation of the actual economic costs of road accidents (now probably underestimated by several tens of percentage points) and the proper priority setting for road safety improvement (Thomas et al, 2003).

2.2 Hazardous Sites Correction Methods: Overview

As explained in the general introduction of this dissertation, typical procedures for hazardous site correction can be divided into three basic tasks (see section 1.3.2):

1. The identification and ranking of hazardous locations. This results in a list of sites with promise ('Identification Phase').

2. Prioritizing these sites by diagnosing the problems at identified locations and determining potential remedial treatments in order to identify cost-effective safety

police, at the international level. Special Report: OECD-RTR programme / Public Roads Administration of Norway. Paris / Oslo.

[6] The author refers to Beaucourt et al.,(1998). Zelfmoord of verkeersongeval

improvement projects ('Investigation Phase').

3. The appraisal of alternative treatments followed by implementation of the best treatment if sufficiently cost effective. To evaluate the effect of treatment, before and after studies need to be conducted ('Program Implementation Phase').

In this section, we will provide an overview of the different methods that are used in literature to perform each of these tasks.

2.2.1 Targeting and Ranking

As described by Vistisen (2002), locations were ranked at first according to their reported number of accidents. Locations with a number exceeding a chosen threshold value were targeted as black spots. However, the author, following Hauer (1986) and Elvik (1997) points out that this method is very sensitive to random variation in accident counts and to the regression to the mean problem (see previous section). Indeed, a site may experience relatively high numbers of crashes due to: (1) an underlying safety problem, for example, the high level of traffic exposure or the nature of the site; or (2) a random 'up' fluctuation in crash counts during the observation period. Simply observing unusually high crash counts does not indicate which of the two conditions prevail at the site (Cheng and Washington, 2005). It is possible to observe an unsafe site that does not reveal elevated crash frequencies, these are termed false negatives. It is also possible to observe elevated crash frequencies at a relatively safe site, these are termed false positives. False positives, if acted upon, lead to investment of public funds with little to no safety benefits. False negatives lead to missed opportunities for effective safety investments. Accordingly, correct determination of hazardous accident locations include identifying a safe site as 'safe' and an unsafe site as 'unsafe'. Therefore, the expected number of accidents, estimated from a model was used instead. However, it is well established that, in general, there are considerable differences between the expected number of accidents at different types of intersections and road sections. This may be inexpedient, as the most effective solution will end up altering the location into a different road type. Such alterations are often too expensive and/or impossible.

Instead of using the number of accidents on a location, McGuigan (1981) suggested ranking sites according to their potential for accident reduction (PAR), which is the difference between the reported number of accidents at a location and the expected number at locations with similar characteristics. In this context, Vistisen (2002) refers to Persaud et al. (1999), who suggested using an empirical Bayes estimate instead of the accident count in PAR. More specifically, they used a Poisson-gamma generalized linear model with characteristics such as traffic flow and various geometric variables. Additionally, in Saccomanno et al (2001) the results of a multivariate Poisson regression model and Empirical Bayes methods are compared for establishing the potential for accidents and designating safety black spots along a highway. The Empirical Bayes model was found to yield fewer black spot locations than the Poisson regression model.

Vistisen (2002) continues that the task of targeting black spots may also be viewed as a ranking and selection problem. Parallel with the PAR-method, Gupta and Hsu (1980) introduced the so-called probability of correct selection (PCS). Then, in a group of locations a subset is targeted as hot spots, if the probability of hereby selecting the site with the largest expected number of accidents (the 'worst' location) is above a chosen threshold value. Later, Hauer and Persaud (1984) derived the probability of correct selection for a Poisson-gamma model. However, the PCS in Hauer and Persaud was used as a measure of the overall efficiency of the targeting method and not directly used for targeting dangerous locations. Schlüter et al (1997) derived the PCS for an individual site as the posterior probability of being 'worst' in a Poisson-gamma model with no location characteristics. Heydecker and Wu (2001) later extended the PCS measure in Schlüter et al to include location characteristics and defined PCS as the probability of the Poisson rate exceeding a chosen threshold.

A few alternative methods for targeting black spots have also been proposed (Vistisen, 2002). For a given treatment measure, Heydecker and Wu (1993) suggested ranking sites according to the posterior probability that accidents occurring at the location involve the feature the measure is aimed at. Heydecker and Wu assumed a Poisson-beta model with no location characteristics. In Persaud and Kazakov (1994) locations are ranked depending on the way in which the estimated economic benefits of treating the location exceed a threshold value based on the allocated budget.

Van den Bossche et al (2002) investigated the question whether a ranking alone can give enough evidence for the selection of dangerous sites. More specifically, Bayesian hierarchical modeling techniques are used to identify and rank hazardous intersections for bicycles in Leuven, a small university town in Belgium. The authors conclude that ranking hazardous sites is an interesting means to get insight in dangerous locations, but there is no such thing as 'the' correct ranking. This finding was confirmed by Miranda-Moreno et al. (2005) who state that many different types of statistical models have been proposed in the past for estimating the accident risk in order to rank locations for safety improvements, ranging from basic Poisson and negative binomial models to more complicated models such as zero-inflated and hierarchical Bayesian models. The authors conclude that the choice of model assumptions and ranking criteria can lead to considerably different lists of black spots. Van den Bossche et al (2002) continue that it should be clear that ranking hazardous sites is an interesting tool to get insight in dangerous locations, but it is by no means an exact enumeration, in ascending order of danger. When investment resources are limited, decision makers are interested in the most hazardous locations. Making a rank order of dangerous sites will undoubtedly be helpful to this end. However, since the ranking is in itself a stochastic term, the obtained rank order may be, to a certain extent, a lucky coincidence. Therefore, other criteria like investment capacity or specific mobility and safety objectives should influence the final ranking.

2.2.2 Prioritizing

As explained in the previous section, Vistisen (2002) describes that high-risk sites are targeted with the aim of improving safety on the road network through remedial treatment of the sites. Any achieved positive effects of safety measures at hazardous accident locations are denoted as the benefits of the implemented measures. However, implementing safety measures is costly and the restricted funding for black spot safety work does put a limit to the number of sites that may be treated. Therefore, it is necessary to prioritize between sites and safety measures in order to utilize the limited funds as effectively as possible.

In this context, Hauer et al. (2004) investigated which of the alternative ranking criteria points to sites with promise at which the most cost-effective solutions can be found. More specifically, they used different ranking criteria ((1) sites where most accidents are expected, (2) sites where most severity-weighed accidents are expected, (3) sites where most excess accidents are expected and (4) sites where most severity-weighed excess accidents are expected) to prepare a list of sites with promise, and made a comparison of the cost-effectiveness of the projects to which these criteria lead. It was found that sites at which most accidents or most severity-weighed accidents are expected lead to most cost-effective projects.

The general aim of prioritizing may be described as:

$$\max_{Y} \frac{B(Y)}{C(Y)} \tag{2.1}$$

where Y represents a portfolio of safety measures and C(Y) and B(Y) denote the corresponding overall cost and benefit of Y.

The objective is to improve safety as much as possible with the funds allocated, but without a given target level of safety. This approach is known as the 'as low as reasonable practicable' (ALARP) principle (Melchers, 2001). Here, 'reasonable practicable' refers to within the budget.

Furthermore, when expressing the benefit and costs of a treatment in objective measures such as saved number of accidents, saved accident costs etc. two important problems can be discerned (Vistisen, 2002). First, the problem of pricing injury accidents is that no market prices are stated for e.g. a human life and instead the pricing of injury accidents is done indirectly (see Dasgupta and Pearce (1972)). There is an extensive body of literature dealing with the valuation of road safety. Although this is beyond the scope of this dissertation, we briefly discuss the two most promininent approaches.

One approach is to assign severity weights that are proportional to the costs to society of various injuries. In general, there are basically three methods for estimating the costs of injury and death to society [7]

1. Using Implicit Values: the accidents are priced according to the average cost of the given medical treatment in trying to avoid a person dying, divided by the probability of the treatment being successful.

2. The method of Human Capital: the major part of the cost of an injury is the discounted present value of the victims future output or income lost due to the injury. The additional cost contributors are involving medical treatment, police, property damage and administration costs.

3. Finally, the Willingness To Pay (WTP) method: this method estimates the value that individuals attach to human life by means of surveys aimed at determining the amount of money that individuals would be prepared to pay to reduce the risk of loss of life. The same principle applies to injury, where an attempt is made to determine the monetary value which individuals would be prepared to

[7] Vistisen refers to T10 (2000), Notes for the unit of the Intercollegiates MSc course in transport. University College London, England.

pay to, in effect, reduce the risk of injury. The advantage of this approach is that it reflects the public‘s concern for safety. Also because WTP values tend to be higher than implicit or human capital values, estimated benefits of remedial work are increased, which may increase the priority given to road safety.

A second approach, originating in public health, is the ‘Quality Adjusted Life Year’ (QUALY) (Nord, 1999). In this approach, any state of illness or disability may be assigned a utility on a scale from zero (the utility assigned to the state of being dead) to unity (the utility assigned to being in full health). The value of a health outcome for an individual is the calculated as the increase in the utility of the persons‘s health state and the number of years the person gets to enjoy this improvement. As argued by Nord (1999), valuations of health based on the person trade-off technique can reasonably be interpreted as representing an ethical point of view with respect to the importance of preventing different adverse health outcomes.

Although they have been developed in different application areas, the QUALY and WTP frameworks share important similarities (Hammitt, 2002): both are justified as representing the preferences of individuals, and both are summed across individuals to represent the social value of a change in health risk. However, QUALYs assume that preferences over health and longevity depend only on health consequences, and do not depend on other characteristics of the individual or the risk. In contrast, WTP allows for the possibility that preferences over health outcomes depend on individual characteristics such as wealth, as well as on characteristics of the risk such as whether it is perceived to be uncontrollable, unfamiliar or dreaded.

In addition to the problem of pricing accidents, the obtained accident reduction from implementing a safety measure needs to be estimated. This can be done using Accident Modification Factors (AMFs). More specifically, AMFs can be used in an accident prediction algorithm to represent the effects on safety of specific geometric design and traffic control features. These AMFs are usually based on a variety of sources including results of before-and-after accident evaluations, coefficients or parameter values from regression models, and expert judgment (Harwood et al., 2000).

However, the obtained accident reduction from implementing a safety measure is often uncertain. This uncertainty may be divided into (Vistisen, 2002):

- The uncertainty concerning the reduction rate in accidents due to the treatment portfolio.

- The uncertainty concerning the extent to which the accident types, the measure is aimed at, are in fact present at the site.

- The uncertainty concerning whether or not a site is in fact a black spot.

The last two uncertainties are site-related, while the first uncertainty is linked to the safety measure. It is assumed that sites with a high certainty of being a black spot have relatively higher potential for accident reduction than sites with a low certainty (see Persaud et al., 1999). Also, a safety measure aimed at a particular type of accident is assumed to have a relatively higher effect at sites where such types are predominant. To increase knowledge in the uncertainty related to the safety measure, additional before and after studies of the effects of treatments are needed.

2.2.3 Before and After Studies

Treating a site may lead to changes in traffic volume, road geometry and other site characteristics thus affecting the accident frequency at the site. However, the literature on before and after studies has pointed out a number of other observable and non-observable elements affecting the change in reported accident count before and after the implementation of preventive safety measures (Vistisen, 2002).

For example, different safety measures may be aimed at the same type of accidents, thus creating an effect overlap if implemented at the same site. If the purpose of the before and after study is to estimate the individual effects of the measures, one needs to account for effect overlap. Additionally, accident treatment measures may, beside positive effects, also have negative effects on safety. In a before and after study of the total number of accidents at a site only the net-effect will show.

Another element is the problem of road user behavioral adjustment. This is the problem of road users adjusting their behavior to a safety measure in such a way that the actual effect of the measure deviates from what was expected. Behavioral adjustment leading to an unchanged level of safety after treatment is known as risk homeostasis (see e.g. Wilde, 1986). This phenomenon is strongly connected to the road users subjective perception of safety.

A statistical effect on before and after studies is the regression to the mean effect. Sites for crash reduction monitoring are selected on the basis of a high number of crashes over a given period. This carries with it a risk that some sites with a low average rate may meet the selection criteria, due to ‘highs’ in the random fluctuations over the given period. It would be expected that the crash rate at such a site would be lower in the period following selection, even in the absence of any improvements, reflecting the true underlying crash rate. The benefits of works at such sites would therefore be overestimated, as some of the apparent reduction attributed to the site treatment would be due merely to the expected regression to the mean (Hauer, 1997).

Additionally, in before and after studies one should take into account the migration effect where accident frequency apparently rises at sites that are untreated but adjacent to treated sites. Vistisen (2002) refers to Boyle and Wright (1984) who proposed a hypothesis that the migration effect was due to a behavioral mechanism based on the idea of road user behavioral adjustment. However, the author adds that the existence of a migration effect has not been verified (see Elvik, 1997), and a study by Maher (1990) indicates that the apparent migration effects may to a large extent be explained by a regression to the mean effect caused by a bias-by-not-selection. In other words, sites are not targeted as black spots because of unusually low (below the expected at sites with similar traits) accident counts. In addition, incorrect coding of the location of accidents as well as changes in traffic flow due to treatment of adjacent sites may be contributing factors.

A next element that is often revealed in a study of time series of accident counts is a trend in the accident development over a period of time. There are many factors influencing general increases or decreases in accident counts. For instance road users are changing their choice of modes and attitude in traffic, for example due to amendments of the law, bigger fines, etc. A before and after study needs to account for the effect of such trends that are otherwise attributed to the treatment.

Furthermore, incomplete reporting of accidents leads to a general underestimation of the road safety problems. A change in the level of reporting at a site will also change its estimated level of safety.

Finally, in Hauer (1997) the central role that prediction plays in the estimation of the safety effect of treatments is discussed. For example, the shortcomings and adaptations of conventional approaches such as the Naïve Before-After study are investigated. In this method the count of ‘before’ period accidents is used to predict what would have been the expected count of after-period accidents had the treatment not been implemented. This way of predicting reflects a naïve and usually unrealistic belief that the passage of time was not associated with changes that affected the safety of the entity under study. Adaptations that are approached concern the factors that are measured and interpreted and the use of a comparison group. Furthermore, the author discusses some approaches that better fit the realities of observational before-after studies. It is shown that the Empirical Bayes approach to estimation not only solves the regression to the mean problem but also yields more precise estimates. Next, estimation will be freed from the constraint of a fixed-duration before period. The key roles played by multi-variable models, which express the safety of an entity as a function of its observable traits, are examined. Finally, the author returns to the central issue of the before and after study, that of estimating what the effect of some intervention is and how this approach has been used to estimate the effect of resurfacing rural roads. Hauer concludes that the subject of how to interpret observational before-after studies has certainly not reached closure. Compared with the number of books on the statistical design and interpretation of experiments, very little is written on observational studies and a great deal needs to be done before a maturity of method and routine of application can be attained or claimed.

2.3 Hazardous Sites Approach in Flanders

2.3.1 Introduction

Statistics show that Flanders, the Flemish speaking community of Belgium, does not set a great example for other European countries in the context of traffic safety. In 2002, on the Flemish roads, 721 persons died and 5,234 persons were seriously injured (National Institute for Statistics, 2002). Consequently, in the 'Mobility Plan Flanders', the Flemish government set itself the objective to halve its arrears compared to the European leaders by 2010. Not only does the Flemish region plan to take measures in the area of enforcement, improved vehicle safety and enhanced road users awareness, it also plans to speed up improvements in infrastructure. For this purpose, over a period of five years, a budget of 100 million EURO per year is reserved. More specifically, to implement these infrastructure changes, the following approach is set up[8]:

- Development of a five year program
- Development of a manual 'Safe Traffic Flanders'
- Collection and analysis of data
- Development of a conceptual solution

[8] This program was developed by the Temporary Consortium Safe Traffic Flanders (3V) and the Administration for Roads and Traffic (AWV).

- Realization and Monitoring

The following paragraphs give a more detailed description of these different steps (Yearbook Traffic Safety, 2003).

2.3.2 Five year Program

First of all, by processing the accident statistics of the National Institute for Statistics (NIS) a list of dangerous accident locations is drawn up. In this process, the following parameters are taken into consideration: the total number of injury accidents and the number of light, serious and fatal injuries. This selection and ranking procedure will be elaborately discussed in third chapter of this dissertation. For now, it suffices to point out that for each location where in the last three years, three or more accidents occurred, a priority score is calculated by summing up the number of injuries at these locations while using weighing values of respectively 1, 3 and 5 for the injury types light, serious and fatal. The accident sites with a priority score of 15 or more, are then considered as dangerous. This procedure results in a list of 1,014 dangerous accident locations. The location with the highest list score receives the highest priority. However, to take into account the social impact of the involvement of vulnerable road users (often cycling students), the accidents involving cyclists are given 50% more weight in the priority list. Furthermore, some dangerous accident locations can not be dissociated from neighboring dangerous sites on the same road. Accordingly, these locations are joint in one cluster and are studied together. However, since one will always take into account the accessibility of an area, this does not necessarily imply that these sites will be redesigned at the same time.

2.3.3 Manual 'Safe Traffic Flanders'

The Manual 'Safe Traffic Flanders' has three objectives:

1. Presenting uniform solutions to enhance traffic safety: This involves the creation of a traffic situation that is uniform in function of the environment and the road function. This new situation will simplify the driving requirements of the road user, which, in turn, will diminish the accident probability.

2. Determining procedures in the design process of a project: This will speed up the realization of corrective measures.

3. Establishing standard solutions and a decision tree: This allows the implementation of previous decisions, while taking into account local limiting conditions and the planning context.

More specifically, the manual is divided into five chapters. Chapter one describes the project card, which is intended to guarantee the uniform collection of the different data that are needed to analyze the accident location. Then, after the basic facts are collected using this project card and the accident site is explored, a first analysis of the accidents will be performed. Next, additional data on the spatial and planning context will be collected in order to propose well-founded solutions for the location. The second chapter provides standard solutions for the following situations: roundabouts, traffic lights regulation, bicycle facilities, pedestrian facilities and additional facilities. Chapter three contains the decision tree and is the most important chapter of the manual. In particular, the choice between the different solutions is made taking into account three different perspectives: the accident analysis, the traffic planning and spatial context (to create a sustainable solution) and finally physical and traffic related limiting conditions. Note that when specifying specific projects, the particular situation and influence of other factors than traffic safety will also need to be taken into account. Then, in chapter four, the different steps and procedures for each project are discussed. Finally, chapter five describes the monitoring system that is necessary to evaluate the effects of the implemented measures.

2.3.4 Collection and Analysis of Data

As explained in the previous paragraph, the project card allows to collect data in a uniform manner. These data are mainly extracted from maneuver diagrams of the accidents and additional spatial information. The analysis of these data is performed using three different perspectives:

1. Analysis of the accident data using the AVOC methodology [9]: This methodology is solely based on the use of data of the registered accidents, supplemented with data on traffic, the road and the environment. Based on the results of this analysis specific infrastructure measures can be implemented to enhance traffic safety.

2. Analysis of the traffic planning and spatial context: In this analysis, the road

[9] AVOC = Aanpak Verkeersongevallenconcentraties, CROW publication 66, 1992

category of the accident location and the spatial context of the site is considered (e.g. built-up area, environmental function,..)

3. Analysis of the traffic data: For most of the dangerous accident locations, the traffic intensity is studied together with the registration of all traffic flows and different vehicle types. These figures are incorporated in the site's traffic flow diagram.

2.3.5 Development of a Conceptual Solution

The development of a conceptual solution occurs step by step by means of a decision tree, as mentioned in the manual. In order to decide on the preferred solution for a site, one starts with the results from the traffic safety analysis. This analysis usually offers a variety of possible measures to diminish the number of accidents at the location. Next, the analysis of the traffic planning and spatial context produces a selection of measures. Indeed, it is not only the objective to find a safe but also a sustainable solution to enhance traffic safety. Additionally, one considers the physical and traffic related limiting conditions to select the appropriate measures. This involves evaluating the capacity of the proposed solution and verifying the physical and environmental issues at the site. Finally, in case still several solutions are appropriate for implementation additional factors such as cost, timing and residual risk are evaluated. This filtering process will eventually lead to one preferred solution.

2.3.6 Realization and Monitoring

After the appropriate measure is chosen, the different steps for the final design and realization of the measure are taken. Once the projects are executed, the effect of measures will be continuously monitored.

2.4 Conclusions

The main problems encountered with the analysis of hazardous accident locations are the random issues in road accidents, accident migration and regression to the mean. To account for these problems, a number of statistical models are described in literature to model the variation in accident frequencies. Furthermore, no clear indication exists of what the best length of a dangerous road segment should be. Finally, the issue of underreporting of road accidents will play a role in the identification of hazardous accident sites.

Next, 'Black Spot Safety Work' can be described as the task of improving road safety through alterations of the geometrical and environmental characteristics of the problematic sites in the existing road network. This work may be divided into three phases.

In the first step, hazardous sites are targeted with the aim of enhancing traffic safety through remedial treatment of the sites. This task may also be viewed as a ranking and selection problem. Researchers have proposed several alternative methods for ranking and targeting black spots. However, there is no such thing as 'the' correct ranking.

Restricted funding does put a limit on the number of sites that may be treated. Therefore, it is necessary to prioritize between sites and safety measures in order to utilize the limited funds as effectively as possible. This is the second step of the black spot safety work. Two important problems can be discerned when calculating the costs and benefits of a treatment. First, no market prices are stated for e.g. a human life and instead the pricing of injury accidents is done indirectly. Secondly, the obtained accident reduction from implementing a safety measure is uncertain.

The third step of black spot safety work involves the realization of before and after studies of the effect of treatment. Important elements when analyzing the change in reported accident counts before and after the implementation of safety measures are overlap in effects, negative effects, road user behavioral adjustment, regression to the mean effect, migration effect, general trends and change in the level of reporting. Accordingly, conventional approaches such as the Naïve Before-After study have a number of shortcomings which could partly be solved with some adaptations. Furthermore, it is shown that the Empirical Bayes approach for estimation not only solves the regression to the mean problem but also yields more precise estimates than the traditional estimation methods. However, very little is written on observational studies and the subject of how to interpret observational before-after studies has certainly not reached closure.

Finally, in this chapter the approach towards hazardous sites in Flanders was discussed. Besides the measures in the area of enforcement, improved vehicle safety and enhance road users awareness, the Flemish government has reserved, over a period of five years, a budget of 100 million EURO per year to speed up improvements in infrastructure. For this purpose, the development of a five year program is set up in which the dangerous accident locations are selected. Furthermore, a manual ‘Safe Traffic Flanders’ will be developed to present uniform solutions, determine procedures in the design process of a project and establish standard solutions to find a safe and sustainable solution to enhance traffic safety.

Chapter 3

Ranking and Selecting Dangerous Accident Locations

In this chapter a sensitivity analysis is performed on the currently used method to rank and select dangerous accident sites in Flanders. Literature has shown that the choice of model assumptions and ranking criteria can lead to considerably different lists of hazardous locations. Accordingly, with this research, we want to sensitize government to carefully choose the criteria for ranking and selecting dangerous accident locations[1].

[1] Parts of this chapter have been published as follows:

- Geurts K., Wets G., Brijs T. Karlis, D. and Vanhoof K. (2006), Ranking and selecting dangerous accident locations: correcting for the number of passengers and batesian ranking plots. Journal of Safety Research 37, 83-91.
- Geurts K., Wets G., Brijs T. and Vanhoof K. (2005) Ranking and selecting dangerous accident locations: Case Study. Urban Transport XI, eds. Brevia and Washra, WIT Press, ISBN: 1-84564-008-X, 229-238.
- Geurts K., Wets G., Brijs T., and Vanhoof K. (2005), Identification and ranking of black spots: sensitivity analysis. Journal of Transportation Research Board,Vol. 1897, pp. 34-42. ISBN: 0-3090-945-X. Also in Electronic Proceedings of the 83th Annual Meeting of the Transportation Research Board, Washington, January 11-15, USA, 17 pp.
- Geurts K. (2005), Geurts K. (2005), Selectie en rangschikking van gevaarlijke punten: een grafische benadering. Jaarboek Verkeersveiligheid 2005, Vlaams Congres Verkeersveiligheid, 2005, Brussel.
- Geurts K., Wets G., Brijs T. and Vanhoof K. (2004), Identifying and ranking dangerous accident locations: Overview Sensitivity analysis. Proceedings of 17th ICTAL Workshop in Tarts, Estonia, 28-30 October, 2004.

3.1 Introduction

The correction of dangerous locations is one avenue that is available to traffic engineers in their endeavor to reduce future accident numbers. Typical procedures for hazardous site correction involve three basis tasks (Schlüter et al., 1997):

1. The identification of hazardous locations
2. A diagnosis of the problems at identified locations and a determination of potential remedial treatments
3. An appraisal of alternative treatments to identify the most cost-effective, followed by implementation of the best treatment (if sufficiently cost-effective).

This chapter focusses on the first stage of these procedures, which commonly involves comparing the accident numbers during some period for all locations, to determine which locations are unusually dangerous. More specifically, we will focus of the currently used method in Flanders to identify hazardous locations. This will be explained in the following section.

3.2 Sensitivity Analysis

3.2.1 Defining 'Dangerous'

As explained in the introduction of this dissertation, there is no universally accepted definition of what should be considered as a dangerous location (Hauer, 1996). Accordingly, different European countries have already been developing different approaches towards the management of black spots. Some examples of definitions of hazardous sites (European Road Federation, 2002):

- the United Kingdom: locations of 300 meters where the sum of the road accidents is higher than twelve in three years.
- the Netherlands: intersections where in a period of three to five years at least ten accidents occurred or where at least five accidents or dangerous situations occurred with some common characteristics.

- Geurts K. (2004), Grote en kleine middelen om de verkeersveiligheid te verhogen: Hoe rangschikken en selecteren we gevaarlijke punten? Jaarboek Verkeersveiligheid 2004, Vlaams Congres Verkeersveiligheid, 2004, Brussel pp.44-46.

- Denmark: road sections or intersections where there has been registered significantly more accidents than could be expected for that type of intersection and same traffic volume. Normally, with this, a minimum criterion of four accidents in five years is used.

- Norway: locations of one hundred meters with more than four casualties.

Whichever criteria is adopted, it is common practice to prepare lists of accident locations, ordered according to their dangerousness. The ordered list is important as locations are generally selected by working down the list until the allocated resources are exhausted for the detailed examination (i.e. the diagnosis and identification of potential treatments), and, perhaps, subsequent treatment of locations (Schlüter et al., 1997).

However, although different definitions exist, in general, locations are classified as hazardous locations after an assessment of the level of risk and the likelihood of a crash occurring at a location. Locations that have an abnormally high number of crashes are then described as crash concentrated, high hazard, hazardous or black spots (Bureau of Transport and Regional Economics of Australia, 2001).

Accordingly, taking these different elements of the different definitions for dangerous accident locations into account, we can conclude that, from a discussion with Prof. dr. Elvik, a fairly precise theoretical definition of a road accident black spot is any location that has (a) a higher expected number of accidents than (b) other similar locations, as a result of (c) specific local risk factors. Therefore, road accident black spots should be defined in terms of the expected number of accidents, not the recorded number, as the recorded number of accidents is greatly influenced by random variation (see section 2.1.2). Moreover, a road accident black spot belongs to a population of similar types of locations, which define the normal level of safety that the black spot presumably deviates from. Finally, the excessive number of accidents at a black spot should be attributable to local risk factors that are amenable to highway engineering treatments. By constructing this theoretical definition, we can evaluate the operational definition of black spots in Flanders in terms of commonly accepted diagnostic criteria.

3.2.2 State of the Art in Flanders

In Flanders, 1,014 accident locations are currently considered as dangerous (Ministry of the Flemish Community, 2001). In order to determine these accident locations, the

Flemish government analyzes the accident data that are obtained from the Belgian Analysis Form for Traffic Accidents. This form should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. Based on these data, each site where in the last three years three or more accidents have occurred, is selected. Then, a site is considered to be dangerous when its score for priority (S), calculated using the following formula, equals 15 or more:

$$S = LI + 3SI + 5DI \tag{3.1}$$

where LI = total number of light injuries
SI = total number of serious injuries
(Each casualty that is admitted more than 24 hours in hospital)
DI = total number of deadly injuries
(Each casualty that died within 30 days after the accident)

Based on equation (3.1), in Flanders currently 1,014 locations are considered as dangerous. To improve the traffic safety on these locations, the Flemish government will each year, starting in 2003 for a period of five years, invest 100 million EURO to redesign the infrastructure of the 800 accident locations with the highest score.

Taking into account the three key elements of the theoretical definition of a black spot (see section 3.2.1), it can be seen that the current definition in Flanders is rather 'simplistic'.

First of all, in the Flemish priority score formula accident counts are used to target locations with a number of accidents exceeding a chosen threshold as dangerous accident locations. This relies on the recorded, rather than the expected number of accidents, which, as explained before, is very sensitive to random variation in accident counts and to the regression to the mean problem (see section 2.2.1).

Furthermore, the definition does not refer to a population of sites with similar traits, hence any location, whether it is a curve, narrow bridge, or a junction, that has recorded three or more accidents will be regarded as a black spot.

However, when comparing the Flemish method for ranking and selecting accident locations with the ranking methods used in other European countries, it can be seen that the Flemish government is one of the few countries that does explicitly consider injury severity by applying differing weighing factors to injuries at different levels of severity (1 for Slight, 3 for serious, 5 for fatal). The choice for the values of these weight parameters used in the priority score formula are mainly a policy decision.

Indeed, on the one hand, weight values can be formulated from an economic point of view. This implies that the values are based on the economic and social costs that are associated with the injuries from road accidents such as medical costs, loss of output, costs of property damage, administrative costs and economic valuation of lost quality of life (Handbook of Road Safety Measures, 2004). On the other hand, weight values can be chosen from an ethical point of view. For example, the use of 1_ 1_ 1 values for respectively a light, serious and fatal injury can seem totally irrational from an economic point of view, while from an ethical viewpoint these values correspond with the prevention of all types of accidents no matter the costs. Therefore, it is up to the government to decide which priorities they want to stress in their traffic safety policy. These priorities can than be translated into different weight values.

Additionally, the calculation of dangerous accident locations in Flanders is currently effected on the level of casualties and not as such on the level of accidents. However, this causes a location to appear more dangerous when more passengers are present in the accidents. Furthermore, a location where four accidents occurred is considered equally dangerous as a location where in the same period of time one accident occurred with four occupants.

Therefore, in this chapter, a sensitivity analysis is performed to investigate the strengths and weaknesses of the currently used method to identify and rank hazardous sites in Flanders. More specifically, we will evaluate how big the impact would be on the current selection of dangerous accident locations when different weight values are used. Additionally, the impact of giving weight to the accident instead of to all the injured occupants of the vehicles is investigated. Furthermore, effects of using the expected number of accidents instead of the historic count data to rank and select accident sites are evaluated. Finally, we will use an optimization framework to automatically generate the 'optimal' weighing values based on the currently used weighing combination. By investigating whether these different ranking criteria have an impact on the selection of dangerous accident locations and estimating how big these effects will be on the current ranking of accident locations ranking criteria, we want to sensitize government to carefully choose the criteria for ranking and selecting dangerous accident locations.

3.2.3 Data

To allow for a sensitivity analysis on the currently used black spot criterion in Flanders, this study is based on the same data used to select and rank the 1,014 currently considered most dangerous accident locations. These data originate from a large data set of traffic accidents obtained from the National Institute of Statistics for the region of Flanders (Belgium) for the period 1997-1999. These data are obtained from the Belgian ‘Analysis Form for Traffic Accidents’ that should be filled out by a police officer for each road accident that occurs on a public road involving casualties. However, since the location of these accidents is only accurately known for the ‘numbered’ roads (i.e. highways, national and provincial roads linking towns) because there is a stone marker every hectometer, the identification of dangerous accident locations is related to roadway segments of numbered roads with a length of 100 meters. Furthermore, each intersection is considered as a possible black spot. Accidents occurring in the direct neighborhood of an intersection (within 50 meters) are also incorporated in the calculations of this intersection. This means that the accident locations that are considered as black spots are either roadway segments of 100 meters or intersections.

These traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident, traffic, environmental conditions, road conditions, human conditions and geographical conditions. The accident data needed to perform this sensitivity analysis will be limited to the number of accidents per accident location. Furthermore, these data will only contain the number of fatalities and the number of light and serious casualties per accident location.

In total, 50,961 traffic accidents with casualties are reported in this period. This results in 23,184 unique accident locations included in the data set. On the one hand, we will focus on the 1,014 most dangerous accident locations to explore the sensitivity of their ranking orders when using adapted weighing values and estimates instead of count values. On the other hand, we will concentrate on all of the 23,184 accident locations and all accident locations where at least 3 accidents occurred between 1997 and 1999 to evaluate the stability of the current list of dangerous accident sites when using different selection criteria.

However, note that in this research , no simulated accident data, as proposed in section 2.1.4, were used. Accordingly, we do not know for certain which sites are truly hazardous and which sites happen to have experienced random up fluctuation in crashes during the period of observation. Therefore, when investigating the impact of different ranking criteria we can only estimate the effects on the current ranking of accident locations and the locations that are currently ranked as most dangerous, without claiming that these sites are indeed the truly most dangerous sites.

3.2.4 Percentage Deviation Value

In order to quantify the effects of changing the ranking and selection criteria of dangerous accident locations, we will use the percentage deviation value (D). This measure allows comparing the elements of two data sets of equal size containing different locations.

$$D = 1 - \frac{G}{T} \tag{3.2}$$

$$\begin{aligned} \text{with } G &= \text{Number of common elements in both data sets} \\ T &= \text{Total number of elements in each data set} \end{aligned}$$

As described in 3.2, the percentage deviation D is calculated by dividing the number of accident locations that do not appear in both data sets by the total number of locations in one data set. When the percentage deviation value is small, the two data sets will contain a great number of common elements. In a traffic safety context, this could for example mean that the current priority list of 800 dangerous locations is rather stable, regardless of the used weighing values or the consideration of the number of passengers in the analysis. Note that the percentage deviation only gives information about the number of locations that do not appear in both ranked data sets and does not take into account internal shifts in the ranking position of these common accident locations.

3.3 Impact of the Weight Values

3.3.1 Alternative Weight Value Combinations

As explained in section 2.2.2, restricted funding for black spot safety work puts a limit to the number of sites that may be treated. Therefore, it is necessary to prioritize between sites and safety measures in order to utilize the limited funds as effectively as possible. Accordingly, it can be argued that the severity of accidents should be taken into account when ranking and selecting dangerous accident locations, as accidents with fatal and serious injuries are more costly in both social and economic terms. The problem, however, is how to find weights that reflect the differences in severity of the different injury types. Again, as explained in section 2.2.2, the problem of pricing injury accidents is that no market prices are stated for e.g. a human life and instead the pricing of injury accidents is done indirectly. Although the different methods to valuate road safety is beyond the scope of this dissertation, the choice for the different weights used in the priority value formula will, however, influence the ranking and selection of the most dangerous accident locations. Therefore, to investigate how big this effect would be on the current ranking of accident sites in Flanders, four different weighing value combinations representing a different attitude towards the traffic safety problem are used:

- 1_ 1_ 1: this combination of weight values assumes that every casualty of a traffic accident is evenly important. Therefore, all accidents are evenly important and should be prevented, regardless of the severity of the injury. As explained in section (3.2.2), this type of weighing scheme represents an ethical rather than an economical point of view to prevent all types of accidents no matter the costs.

- 1_ 1_ 10: using these weight values, attention will be focused on accidents with deadly injured casualties. Accidents with lightly or seriously injured casualties receive relatively small attention. This type of weighing scheme could represent the objective of the Belgian government to reduce the number of road fatalities by the year 2010 by 50 percent (see section (1.1)).

- 1_ 10_ 10: this group of weight value combinations discriminates between accidents with small injuries on the one hand and accidents with serious or deadly injuries on the other hand. Lightly injured persons, are assumed to be characteristic for less serious accidents and will be less taken into account when identifying hazardous sites. As explained before, this type of weighing scheme corresponds with the philosophy the ‘Vision Zero’, which is an expression of the ethical imperative that it can never be ethically acceptable that people are killed or seriously injured when moving within the road transport system (Gingival and Tamworth, 1999).

- 1_ 3_ 5: These combinations of weight values use a more moderate approach to stress the importance of deadly accidents. As the injury types are more serious, the accident is considered to be more important. Note that the current dangerous accident selection criterion in Flanders is based on this proportion of weighting values. However, no justification on these proportions can be found in any public document concerning this policy.

These weighing combinations show that different policies can be used to arrive at a weighing scheme. However, the choice for using, in this research, a value of '10' to stress the importance of an injury type is somewhat arbitrary. Nevertheless, it is not the objective of this sensitivity analysis to point out how one should best arrive at a weighing scheme. This research is focussed on the first and necessary step to make the government aware that choosing different weighing values will greatly influence the selection of hazardous accident locations. This might seem straightforward, however, in Flanders no such research is yet conducted. Accordingly, the weighing values used in this sensitivity analysis are just examples and therefore no statement is made that one of these four alternatives is the optimal weighing scheme.

In the following sections, we will discuss the effects of using these alternative weighting value combinations on the 3 different data sets: the 1,014 currently most dangerous accident locations, all 23,184 accident locations and the 5,326 accident locations with minimum 3 accidents.

3.3.2 Empirical Results

3.3.2.1 Results for the 1,014 Currently Most Dangerous Accident Locations

Figure 3.1 gives an overview of the effects of changing the 1_ 3_ 5 weight values on the ranking order of the 1,014 accident locations that are currently considered as most dangerous. In particular, figure 3.1 shows the percentage deviation values for different subsets of the 1,014 currently dangerous accident locations, using respectively the weighing values 1_ 1_ 1, 1_ 1_ 10 and 1_ 10_ 10.

Additionally, as explained in section (3.2.2), for budgetary reasons, the government does not redesign all 1,014 dangerous accident locations, but only the 800 accident sites with the highest priority score. Since different weighing value combinations produce different ranking orders, table 3.1 shows the percentage deviation values for the top 15%, top 40%, top 70% and top 800 of the 1,014 currently most dangerous accident locations using different combinations of weighing values.

These results show that when changing the 1_ 3_ 5 weight value combinations, this causes the different location subsets to deviate from the original 1_ 3_ 5 location subsets from 12.5% to 23.5%. This means that in the most extreme case, 23.5% of the accident locations considered belonging to the 15% most dangerous accident locations of the 1,014 dangerous accident locations using 1_ 3_ 5 weight values do not appear

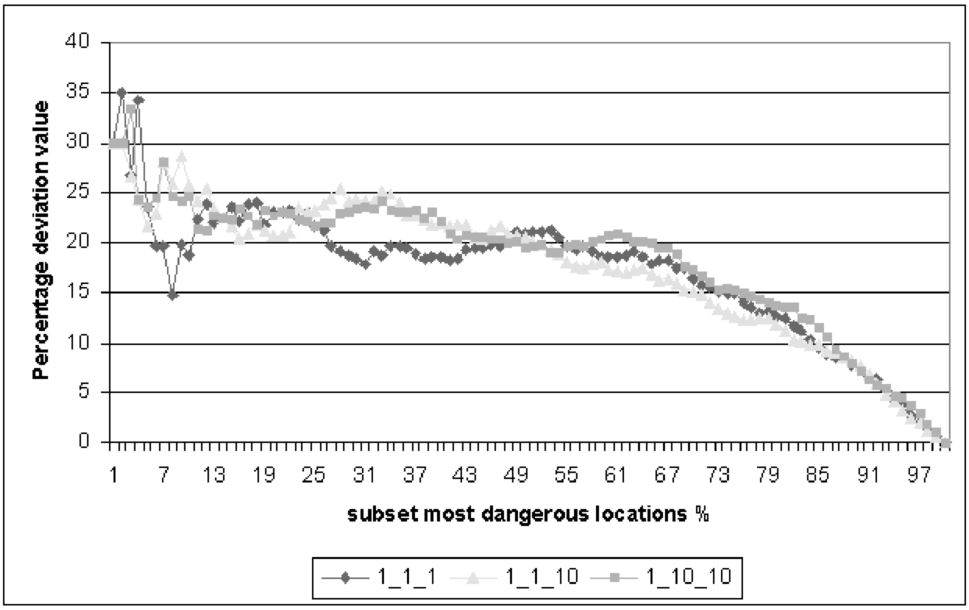

Figure 3.1: Percentage deviation values for the 1,014 currently most dangerous accident sites using alternative weighing value combinations than 1_ 3_ 5

in the top 15% when changing the weight values to 1_ 1_ 1. On the other hand, when changing the weight values to the combination 1_ 1_ 10, only 12.5% of the resulting most dangerous accident locations will differ from the currently 800 selected black spots.

Furthermore, results of table 3.1 show that the percentage deviation values of the different weighing value combinations differ greatly amongst each other. The more the weight values deviate from each other, the greater the variability of the black spots will be. For example, when comparing the top 15% (of the 1,014 most dangerous accident locations) calculated using the weight value combination 1_ 10_ 10 and the weight value combination 1_ 1_ 1, table 3.1 shows that 43.8% of the accident locations in this subset will differ from each other. When comparing this result with the previous results, it can be seen that this percentage deviation value is more extreme than the values related to the 1_ 3_ 5 weight combination (respectively 23.5% and 22.8% for the top 15%). This can be explained by the more 'general' character of the 1_ 3_ 5 weight values, which does not differ as greatly from the other weight value combinations. However, note that although both weight value combinations

Table 3.1: Percentage deviation values for different subsets of the 1,014 currently most dangerous accident locations using different combinations of weight values

	X (top X)	LI_ SI_ DI			
LI_ SI_ DI		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
1_ 1_ 1	15%	0%	_	_	_
1_ 1_ 1	40%	0%	_	_	_
1_ 1_ 1	70%	0%	_	_	_
1_ 1_ 1	800	0%	_	_	_
1_ 1_ 10	15%	30.0%	0%	_	_
1_ 1_ 10	40%	26.1%	0%	_	_
1_ 1_ 10	70%	16.7%	0%	_	_
1_ 1_ 10	800	11.8%	0%	_	_
1_ 3_ 5	15%	23.5%	21.5%	0%	_
1_ 3_ 5	40%	18.4%	21.9%	0%	_
1_ 3_ 5	70%	16.4%	15.2%	0%	_
1_ 3_ 5	800	13.1%	12.5%	0%	_
1_ 10_ 10	15%	43.8%	40.5%	22.8%	0%
1_ 10_ 10	40%	39.4%	39.6%	22.1%	0%
1_ 10_ 10	70%	33.1%	40.5%	22.8%	0%
1_ 10_ 10	800	23.4%	23.7%	14.1%	0%

practically have an equal impact on the number of accident locations that will differ in the top 15%, the actual accident locations that change in this subset will not be the same.

Finally, the percentage deviation values are on average smaller when more accident locations are involved in the analysis. A possible explanation could be that the accident locations with a higher ranking value (i.e. the most dangerous sites), according to the 1_ 3_ 5 weight values, are more sensitive to changes in the weight values than the accident locations with smaller ranking values. However, we should take into account that the larger the top of accident locations that is considered for analysis (X), the more accident locations can obtain a different ranking order without falling out of the top X% most dangerous accident locations. Indeed, as explained in one of the previous sections, the percentage deviation value gives no information about these internal changes in ranking orders. More specifically, this means that, when

Table 3.2: Total number of fatal, serious and light injuries for the 1,014 currently most dangerous sites

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	249	2,326	13,246	15,821
1_ 1_ 10	374	2,267	12,814	15,455
1_ 3_ 5	329	2,571	12,496	15,396
1_ 10_ 10	341	2,817	11,280	14,438

using different weighing values, selecting the accident sites belonging to the top 800 and the top 70% (710 sites) of the most dangerous accident locations, the maximum number of sites that can differ from the original 1_ 3_ 5 selection is respectively 26.75% ((1,014 - 800)/800) and 42.8% ((1,014 - 710)/710). For the top 15% (152 sites) and the top 40% (406 sites), 100% of the selected sites can differ from the original 1_ 3_ 5 selection. Accordingly, we should not compare the absolute values of the percentage deviation values amongst different subset sizes (vertically) but only amongst different weighing value combinations selecting the same subset size (horizontally).

Additionally, table 3.2 shows that using the current 1_ 3_ 5 weights, selecting and tackling the 800 most dangerous accident locations relates to 329 deadly injured victims (DI), 2,571 seriously injured persons (SI) and 12,496 lightly injured persons (LI). Depending on the chosen alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 249 to 374), seriously injured persons (ranging from 2,267 to 2,817) and lightly injured persons (ranging from 11,280 to 13,246). The total number of injured persons will then vary from 14,438 to 15,821. As explained in the beginning of this chapter (see section 3.3.1), this choice between avoiding as much victims as possible or prioritizing more serious injury accidents represent different attitudes from the government towards the traffic safety problem.

These results illustrate that a different attitude towards the traffic safety problem and the choice of the corresponding injury weight values in combination with the size of the subset of the accident locations that is considered has important consequences for the selection and ranking of hazardous sites and for traffic safety actions in general.

3.3.2.2 Results for all 23,184 Accident Locations

In this section, we look at the consequences of changing the injury weight values on all 23,184 accident locations of Flanders. This allows to explore the sensitivity of the ranking orders of all accident locations, including the locations that are currently not considered as being very dangerous.

Figure 3.2 gives an overview of the effects of changing the 1_ 3_ 5 weight values on the ranking order of all 23,184 accident locations that are currently considered as most dangerous. In particular, figure 3.2 shows the percentage deviation values for the three combinations of weight values representing different attitudes towards the traffic safety problem (see previous section) for different subsets of all accident locations.

In addition, table 3.3 shows the percentage deviation values for the top 800, top 15%, top 40%, top 70% of all 23,184 accident locations using different combinations of weighing values.

Results of this table show that when selecting the 800 most dangerous accident locations using the different combinations of weight values, the 'new black spots'

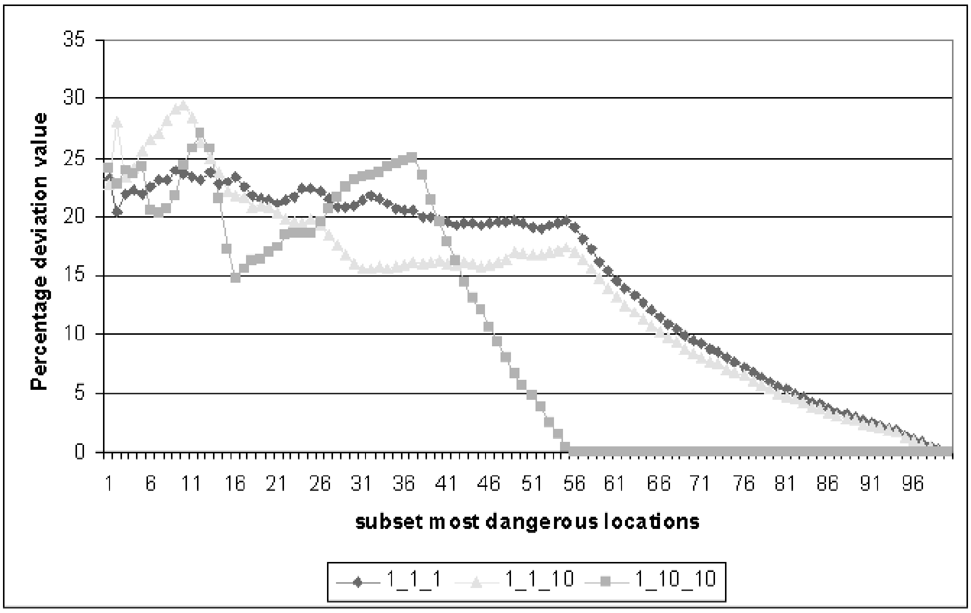

Figure 3.2: Percentage deviation values for all accident sites using alternative weighing value combinations than 1_ 3_ 5

Table 3.3: Percentage deviation values for different subsets of all accident locations using different combinations of weight values

	X (top X)	LI_ SI_ DI			
LI_ SI_ DI		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
1_ 1_ 1	800	0%	_	_	_
1_ 1_ 1	15%	0%	_	_	_
1_ 1_ 1	40%	0%	_	_	_
1_ 1_ 1	70%	0%	_	_	_
1_ 1_ 10	800	34.0%	0%	_	_
1_ 1_ 10	15%	25.6%	0%	_	_
1_ 1_ 10	40%	5.9%	0%	_	_
1_ 1_ 10	70%	3.1%	0%	_	_
1_ 3_ 5	800	21.6%	24.6%	0%	_
1_ 3_ 5	15%	22.8%	22.3%	0%	_
1_ 3_ 5	40%	19.9%	15.8%	0%	_
1_ 3_ 5	70%	9.8%	8.7%	0%	_
1_ 10_ 10	800	44.6%	44.3%	24.5%	0%
1_ 10_ 10	15%	38.5%	36.4%	17.1%	0%
1_ 10_ 10	40%	39.4%	35.4%	19.5%	0%
1_ 10_ 10	70%	33.1%	8.9%	0.5%	0%

will contain between 21.6% and 24.6% different accident locations compared to the 800 most dangerous accident locations selected by the 1_ 3_ 5 values. These results show that the concept of dangerous accident locations is rather relative and depends strongly on the chosen traffic safety policy. Note that these figures differ from the results in table 3.1 for the top 800. This can be explained by the fact that the top 800, which in both tables is selected by the 1_ 3_ 5 conditions and which is used to calculate the percentage deviation values with the other weight values, differs in the two tables. In this last analysis, we do not select the locations with a minimum of 3 accidents in the last 3 years first. Nor does the priority value need to exceed 15.

Furthermore, when considering subsets of different sizes of all the accident locations and comparing the results with these of the 1_ 3_ 5 weight values the percentage deviation values will vary from 0.5% to 24.6%. Similar to the results of the previous section, these percentage deviation values are on average smaller when more accident locations are involved in the analysis. For example, when using the 1_ 10_ 10 values to determine the subset of 70% most dangerous accident locations, only 0.5% of these accident locations will differ from the results of the 1_ 3_ 5 weight combination. As explained in the previous section, this can be a result from the inclusion of more accident locations which allows more internal shifts without falling out of the subset. More specifically, for this data set this means that, when using different weighing values, selecting the accident sites belonging to the top 70% (16,229 sites) of the most dangerous accident locations, the maximum number of sites that can differ from the original 1_ 3_ 5 selection is 42.85% ((23,184 - 16,229)/16,229). For the top 800, the top 15% (3,478 sites) and the top 40% (9,274 sites), 100% of the selected sites can differ from the original 1_ 3_ 5 selection. Accordingly, we should not compare the absolute values of the percentage deviation values amongst different subset sizes (vertically) but only amongst different weighing value combinations selecting the same subset size (horizontally). However, when comparing this percentage deviation value of 0.5% for the 1_ 10_ 10 values with the the percentage deviation value for the 1_ 1_ 1 weight values to identify the top 70% (9.8%), we can conclude that, although the size of the subset is large, still some significant changes can appear depending on the weighing values.

Next, comparing the differences in the selection of the accident locations by the different weight value combinations amongst each other leads to some interesting results in accordance with the results of table 3.1. For example, when comparing the top 800 of all accident locations calculated using the weight value combination 1_ 10_ 10 and the weight value combination 1_ 1_ 1, table 3.3 shows that 44.6% of the accident locations in this subset differ from each other. When comparing this result with the results related to the 1_ 3_ 5 weight values again it can be seen that this percentage deviation value is more extreme for the different weight value combinations amongst each other. Additionally, although results from table 3.3 show that the impact of both weight value combinations on the number of accident locations that will differ in the top 800 does not vary strongly (respectively 21.6% and 24.5% for the 1_ 1_ 1 and 1_ 10_ 10 combination), the actual accident locations that change in this subset will not be the same.

Additionally, table 3.4 shows that using the current 1_ 3_ 5 weights, selecting and tackling the 800 most dangerous accident sites relates to 401 deadly injured victims (DI), 2,636 seriously injured persons (SI) and 12,141 lightly injured persons (LI). When using an alternative weighing combination, the selection of the 800 most

Table 3.4: Total number of fatal, serious and light injuries for all 23,184 accident sites

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	234	2,144	13,557	15,935
1_ 1_ 10	635	2,006	11,554	14,195
1_ 3_ 5	401	2,636	12,141	15,178
1_ 10_ 10	399	3,112	9,727	13,238

dangerous sites will cover a different number of respectively fatally injured persons (ranging from 234 to 635), seriously injured persons (ranging from 2,006 to 3,112) and lightly injured persons (ranging from 9,727 to 13,557). The total number of injured persons will then vary from 13,238 to 15,935.

3.3.2.3 Results for the 5,326 Accident Locations with Minimum 3 Accidents

In this last section, we will investigate the effects of changing the injury weight values on all accident locations of Flanders where at least 3 accidents occurred between 1997 and 1999. Note that this is currently one of the conditions (besides the priority score) for an accident location to be considered as dangerous. Therefore, analyzing the impact of the injury weight values on the ranking orders of these 5,326 accident locations allows a more realistic exploration of the sensitivity of the ranking and selecting of dangerous accident locations when using different weighing values.

Figure 3.3 gives an overview of the effects of changing the 1_ 3_ 5 weight values on the ranking order of the accident sites with minimum 3 accidents. In particular, figure 3.3 shows the percentage deviation values for different subsets of these 5,326 accident locations, using respectively the weighing values 1_ 1_ 1, 1_ 1_ 10 and 1_ 10_ 10..

In addition, table 3.5 shows the percentage deviation values for the top 800, top 15%, top 40%, top 70% of all 23,184 accident locations using different combinations of weighing values.

More specifically, table 3.5 shows the percentage deviation values for different combinations of weighing values and for different subsets of the accident locations with at least 3 accidents, all related to the ranking order of the currently used 1_ 3_ 5 weighing values. Note that for these locations the top 15% coincides with the top 800.

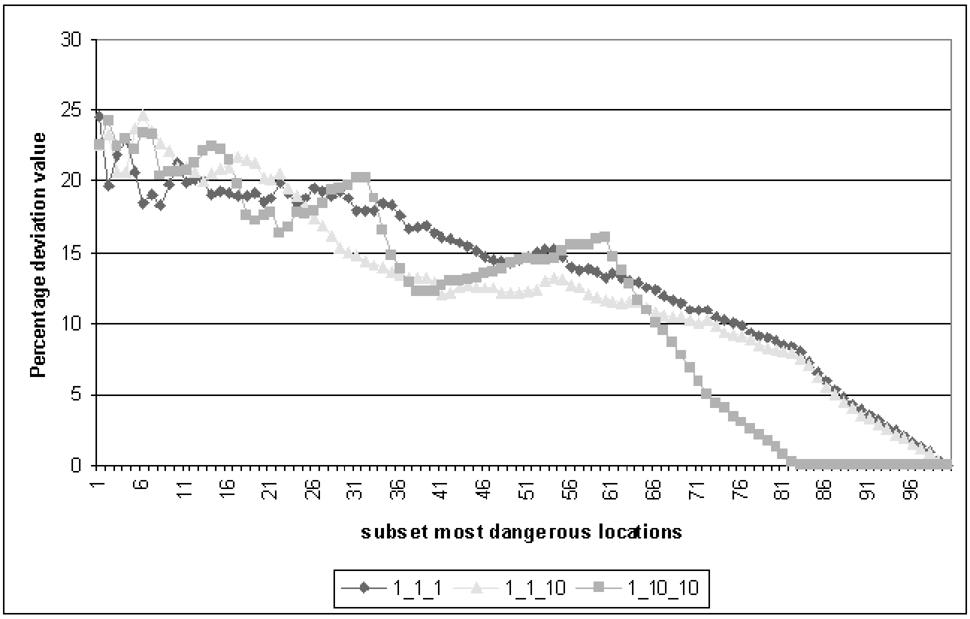

Figure 3.3: Percentage deviation values for accident sites with minimum 3 accidents using alternative weighing value combinations than 1_ 3_ 5

These results show that depending on the chosen injury weight values, selecting subsets of the most dangerous accident locations can deviate from the 1_ 3_ 5 location subsets from 6.9% to 22.1%. More specifically, when selecting the 800 most dangerous accident locations using the 1_ 10_ 10 weight values, 22.1% of these accident locations will differ from the 1_ 3_ 5 selection. Note that for the top 800, these results strongly coincide with the results of table 3 although in this analysis the extra criterion of minimum 3 accidents per location was used. Furthermore, analogously to the results of table 3.1 and table 3.3, the percentage deviation values are on average smaller when more accident locations are involved in the analysis.

More specifically, for this data set this means that, when using different weighing values, selecting the accident sites belonging to the top 70% (3,729 sites) of the most dangerous accident locations, the maximum number of sites that can differ from the original 1_ 3_ 5 selection is 42.83% ((5,326 - 3,729)/3,729). For the top 800 (15%) and the top 40% (2,131 sites), 100% of the selected sites can differ from the original 1_ 3_ 5 selection. Accordingly, we should not compare the absolute values of the percentage deviation values amongst different subset sizes (vertically) but only amongst different

Table 3.5: Percentage deviation values for different subsets of the accident locations with minimum 3 accidents using different combinations of weight values

	X (top X)	LI_ SI_ DI			
LI_ SI_ DI		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
1_ 1_ 1	800 (15%)	0%	_	_	_
1_ 1_ 1	40%	0%	_	_	_
1_ 1_ 1	70%	0%	_	_	_
1_ 1_ 10	800 (15%)	27.5%	0%	_	_
1_ 1_ 10	40%	9.0%	0%	_	_
1_ 1_ 10	70%	1.8%	0%	_	_
1_ 3_ 5	800 (15%)	19.2%	20.9%	0%	_
1_ 3_ 5	40%	16.4%	12.8%	0%	_
1_ 3_ 5	70%	10.9%	10.2%	0%	_
1_ 10_ 10	800 (15%)	40.0%	34.6%	22.1%	0%
1_ 10_ 10	40%	26.1%	24.5%	12.2%	0%
1_ 10_ 10	70%	17.9%	17.1%	6.9%	0%

weighing value combinations selecting the same subset size (horizontally).

Additionally, table 3.6 shows that when using the current 1_ 3_ 5 weighing combination, selecting and tackling the 800 most dangerous accident sites relates to 329 deadly injured victims (DI), 2,571 seriously injured persons (SI) and 12,496 lightly injured persons (LI). When using an alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 223 to 481), seriously injured persons (ranging from

Table 3.6: Total number of fatal, serious and light injuries for the sites with minimum 3 accidents

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	223	2,126	13,580	15,929
1_ 1_ 10	481	2,066	12,254	14,801
1_ 3_ 5	329	2,571	12,496	15,396
1_ 10_ 10	338	2,943	10,457	13,738

2,006 to 2,943) and lightly injured persons (ranging from 10,457 to 13,580). The total number of injured persons will then vary from 13,738 to 15,929.

3.3.3 Conclusions

In this section, we used four different weighing value combinations each representing a different attitude towards the traffic safety problem to investigate how big the effect would be of using one of these alternative policies on the current selection of hazardous accident sites in Flanders. More specifically, we discussed the effects of using the weighing value combinations 1_ 1_ 1, 1_ 1_ 10, 1_ 3_ 5 and 1_ 10_ 10 (for respectively each deadly injured, seriously injured and lightly injured person), on three different data sets: the 1,014 currently most dangerous accident locations, all 23,184 accident locations and the 5,326 accident locations with minimum 3 accidents.

Results showed that a change in the traffic safety policy and the reflection of this choice in the injury weighing values used to rank and select the most dangerous accident locations will not only have an important impact on the number of accident sites that will change when selecting the most hazardous sites, it will also have an important effect on the type of accident locations (e.g. locations with high traffic volumes resulting in many small accidents with mostly light injuries opposite to locations with lower traffic volumes but severe injuries) that are selected. More specifically, depending on the chosen alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured, seriously injured and lightly injured persons. This choice between avoiding as much victims as possible or prioritizing more serious injury accidents represent different attitudes from the government towards the traffic safety problem and accordingly will have an impact on the resulting future traffic safety decisions. Government should therefore carefully decide which priorities should be stressed in the traffic safety policy. Then, the according weighing value combination can be chosen to rank and select the most dangerous accident locations.

Note that, as explained in section (3.3) the weighing values used in this sensitivity analysis are just examples and therefore no statement is made that one of these four alternatives is the optimal weighing scheme .

3.4 Impact of the Number of Passengers

3.4.1 Weighing the Severity of the Accident

In this section, we will evaluate how big the effect would be on the current ranking of the accident locations in Flanders when giving a weight to the severity of the accident instead of to all the injured occupants of the vehicles. This choice is motivated by preliminary results, which showed that most of the 1,014 accident locations that are currently considered as dangerous on average count 10 accidents. In other words, it turns out that not the number of accidents but the number of injured passengers is the most differentiating factor for the ranking of these accident locations. However, if we assume the number of passengers in the car to be a random variable with equal mean across all accident locations, then the number of passengers involved in the accidents on a particular location should not reflect the scoring of that location. We will therefore calculate the scoring of each location based on the severity of the accident and not on the severity of the involved casualties, the former which is free of any bias resulting from coincidental circumstances about the number of passengers in the car. In practice, this means that the points per accident that are summed up in order to calculate the priority value of the locations can vary between 1 (only light injuries), 3 (at most serious injuries) and 5 (at least one deadly injured casualty).

3.4.2 Empirical Results

Figure 3.4 gives an overview of the effects of giving weight to the severity of the accident instead of to all the injured occupants of the vehicle. In particular, figure 3.4 shows the percentage deviation values for different subsets of the three data sets (1,014 currently dangerous accident locations, locations with minimum 3 accidents and all accident locations,) only taking into account the most serious injury per accident. These results show that correcting for the number of passengers will indeed cause a change in the ranking and selection of the most dangerous accident locations.

In addition, table 3.7 shows the percentage deviation values for the top 15%, top 40%, top 70% and top 800 of the most dangerous accident sites when correcting for the number of passengers.

More specifically, when looking at the first data set, the 1,014 accident locations that are currently considered as dangerous, giving weight to the accidents instead of to all the injured passengers, while keeping the 1_ 3_ 5 weighing values, causes the different location subsets to deviate from the original location subsets up to 25.0%

for the top 15% and 14.7% for the top 800. This means that 14.7% of the accident locations that are currently considered to belong to the 800 most dangerous accident locations do not appear in the top 800 when correcting for the number of passengers. When using different weighing values, the comparison between correcting for the number of passengers or counting all the injured passengers also produces percentage deviation values of the same magnitude (see table 3.7).

When looking at the data set containing all of the 23,184 accident locations, correcting for the number of passengers when using the 1_ 3_ 5 weighing values results in a maximum percentage deviation value of 24.2% (top 800). This means that 24.2% of the accident locations that are currently considered to belong to the 800 most dangerous accident locations do not appear in the top 800 when only taking into account the most serious injury per accident.

Finally, results for the 5,326 accident locations with minimum 3 accidents indicate

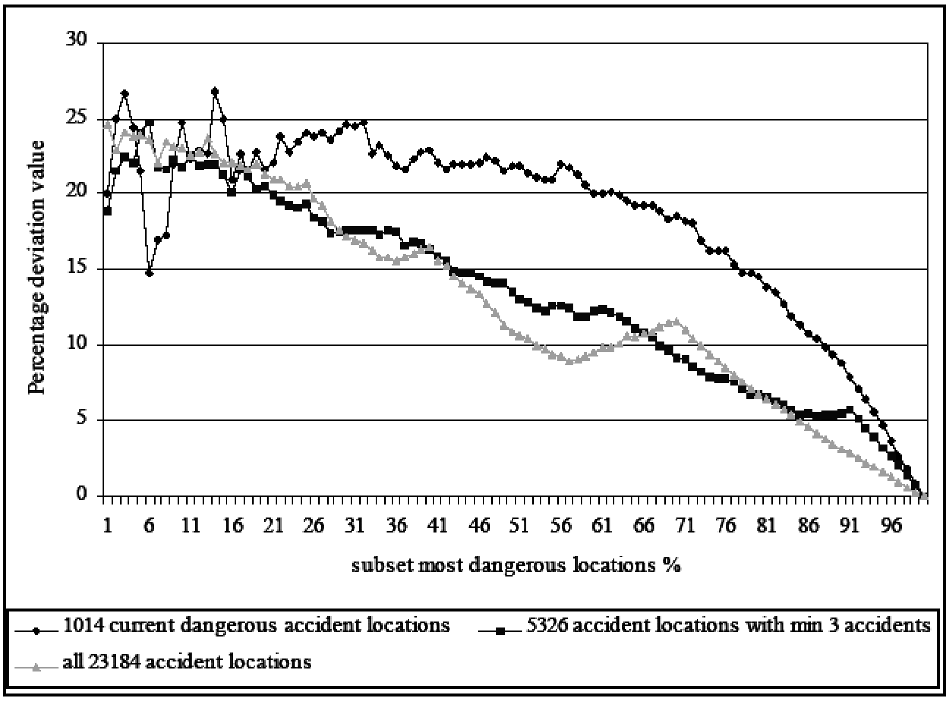

Figure 3.4: Percentage deviation values for different subsets of the accident locations taking into account the most serious injury per accident

Table 3.7: Percentage deviation values for correcting for the number of passengers versus not correcting

1,014 currently most dangerous sites	top 15%	top 40%	top 70%	top 800
1_ 1_ 1	24.3%	22.4%	12.9%	9.1%
1_ 1_ 10	24.3%	17.9%	12.1%	9.2%
1_ 3_ 5	25.0%	22.9%	18.6%	14.7%
1_ 10_ 10	31.4%	23.6%	14.1%	9.7%

all 23,184 accident locations	top 800	top 15%	top 40%	top 70%
1_ 1_ 1	23.2%	23.1%	20.9%	13.5%
1_ 1_ 10	28.3%	14.5%	18.9%	14.9%
1_ 3_ 5	24.2%	22.1%	16.6%	12.2%
1_ 10_ 10	31.5%	26.6%	3.1%	12.3%

locations with minimum 3 accidents		top 800 (15%)	top 40%	top 70%
1_ 1_ 1		24.6%	20.1%	15.4%
1_ 1_ 10		19.8%	16.8%	13.6%
1_ 3_ 5		21.4%	16.2%	9.1%
1_ 10_ 10		25.5%	11.6%	5.2%

that, using the 1_ 3_ 5 values, giving weight to the accidents causes the different location subsets to deviate from the original location subsets up to 21.4% (top 800 or 15%).

Analogously with the results for the 1,014 currently most dangerous accident locations, for all the 23,184 accident sites and the 5,326 locations with minimum 3 accidents, when using different weighing values, correcting for the number of passengers produces percentage deviation values of the same magnitude.

Note that again, for the three data sets the percentage deviation values are on average smaller when more accident locations are selected in top for analysis.

Additionally, table 3.8 shows that selecting the top 15% or 799 most dangerous accident sites from the 1,014 currently most dangerous accident locations, using the 1_ 3_ 5 weights and correcting for the number of passengers, relates to 276 deadly injured victims (DI), 1,819 seriously injured persons (SI) and 7,178 lightly injured persons (LI). Additionally, when not only correcting for the number of passengers but

Table 3.8: Total number of fatal, serious and light injuries for the 1,014 currently most dangerous sites when correcting for the number of passengers

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	213	1,727	7,526	9,466
1_ 1_ 10	318	1,634	7,291	9,243
1_ 3_ 5	276	1,819	7,178	9,273
1_ 10_ 10	288	1,934	6,512	8,734

also choosing alternative weighing value combinations, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 213 to 318), seriously injured persons (ranging from 1,634 to 1,934) and lightly injured persons (ranging from 6,512 to 7,526). The total number of injured persons will then vary from 8,734 to 9,466. Compared with the results of table 3.2, for all the weighing value combinations, the total number of injured persons that will be selected is smaller when correcting for the number of passengers. These results are exactly the consequence of calculating the scoring of each location based on the severity of the accident and not on the severity of the involved casualties, the former which is free of any bias resulting from coincidental circumstances about the number of passengers in the car.

Similar results can be found for the data set with all the 23,184 accident locations. More specifically, selecting the 800 most dangerous sites using the 1_ 3_ 5 weights and correcting for the number of passengers, relates to 272 deadly injured victims (DI), 1,822 seriously injured persons (SI) and 7,294 lightly injured persons (LI). Additionally, when not only correcting for the number of passengers but also choosing

Table 3.9: Total number of fatal, serious and light injuries for all 23,184 accident sites when correcting for the number of passengers

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	165	1,473	8,202	9,840
1_ 1_ 10	611	1,176	5,885	7,672
1_ 3_ 5	272	1,822	7,294	9,388
1_ 10_ 10	249	2,081	6,206	8,536

Table 3.10: Total number of fatal, serious and light injuries for the sites with minimum 3 accidents when correcting for the number of passengers

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	165	1,469	8,198	9,832
1_ 1_ 10	531	1,277	6,470	8,287
1_ 3_ 5	276	1,820	7,269	9,365
1_ 10_ 10	248	2,081	6,192	8,521

alternative weighing value combinations, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 165 to 611), seriously injured persons (ranging from 1,176 to 2,081) and lightly injured persons (ranging from 5,885 to 8,202). The total number of injured persons will then vary from 7,672 to 9,840.

Finally, selecting the 800 most dangerous locations from the 5,326 accident sites with minimum 3 accidents,using the 1_ 3_ 5 weights and giving weight to the severity of the accidents, results in the selection of 276 deadly injured victims (DI), 1,820 seriously injured persons (SI) and 7,269 lightly injured persons (LI). When also using different weighing value combinations, the selection of the 800 most dangerous sites will cover between 165 and 531 fatally injured persons, between 1,277 and 2,081 seriously injured persons and between 6,192 and 8,198 lightly injured persons. The total number of injured persons will then vary from 8,287 to 9,840.

3.4.3 Conclusions

This research showed that giving weight to the severity of the accident instead of to all the injured occupants of the vehicle will have a big impact on the selection and ranking of dangerous accident locations. More specifically, when selecting the 800 most dangerous accident sites of all accident locations, when only taking into account the most serous injury per accident, 24.2% of these locations will differ from the current selection. Accordingly, giving weight to the severity of the accident corrects for the bias that occurs when the number of occupants of the vehicles are subject to coincidence. Obviously, this correction results in different accident sites that are selected as most dangerous, which will also have an important effect on the resulting future traffic safety policy and decisions. Furthermore, the total number of injured persons that will be selected are smaller when correcting for the number of passengers. These results are exactly the consequence of calculating the scoring of each location based on the severity of the accident and not on the severity of the involved casualties, the former which is free of any bias resulting from coincidental circumstances about the number of passengers in the car. However, in some cases (e.g. discotheques, entertainment centers), it can be reasoned that the number of occupants, and accordingly the number of injured persons, is not a coincidence but more likely a trend. For these locations, correcting for the number of passengers would not be advisable since the number of injuries that appear at these locations are inherent to the locations characteristics.

3.5 Impact of Using the Expected Number of Accidents

3.5.1 Estimating the Expected Number of Accidents

Literature points out that when identifying hazardous locations, it is common practice to prepare lists of accident locations, ordered according to their empirical accident rate (Schlüter et al., 1997). However, as explained by Hauer (1986) the actual count of accidents are subject to random variation and to the regression to the mean problem. Accordingly, locations that in one period recorded 'x' accidents do not have, on the average, 'x' accidents in the subsequent period. Therefore, the actual count of accidents is not a very good estimate for the expected number of accidents at a location. Consequently, ordered lists constructed by ranking locations according to their empirical accident rate, and thus ignoring the variability associated with each estimate, do not ensure that the worst locations will be identified (Schlüter et al., 1997).

To account for this random character of accident counts, a number of statistical models have been used to estimate the accident frequency at a specific location over a given interval of time. More specifically, compelling arguments can be found to support the assumption that accident counts follow the Poisson probability law with parameter λ (Nag et al., 2002). The underlying assumption is that road accidents can be treated as random events with an underlying mean accidents rate for each accident location.

Formula (3.3) shows a Poisson distribution where x_i represents the actual number of accidents ($x_i = 0, 1, 2,..$) and λ_i the expected number of accidents for a specific location:

$$f(x_i \mid \lambda_i) = \frac{e^{-\lambda_i}\ \lambda_i^{x_i}}{x_i!} \tag{3.3}$$

The expected value λ_i of a location is the unknown parameter that needs to be estimated so that we can use this value to rank the locations.

In traditional statistical approaches, this parameter will be estimated using the frequency theory, which defines the probability of a characteristic as its long term frequency. For example, suppose we want to estimate the probability of an event A where N represents the number of times the experiment has been carried out and M the number of times the event A occurred. Equation (3.4) shows that the long term frequency of this event A will then be estimated using the limit of the relative frequency, introduced by Denis Poisson (Jackman, 2002):

$$Pr(A) = \lim_{n \to \infty} \frac{M}{N} \tag{3.4}$$

From this equation it should be clear that when it is not possible to carry out repetitive experiments, or when little data are available or the sample size is small, traditional statistical approaches are not very suitable to estimate the probability of an event. This is often the case in traffic research, where the availability of data is usually sparse with short periods of observation. As a result, Bayesian methods have received increasing support from traffic researchers attempting to overcome these difficulties (Schlüter et al., 1997).

3.5.2 Bayesian Statistical Models

3.5.2.1 Bayesian Models in Traffic Safety Research

The interest in the use of Bayesian estimation methods in the domain of traffic safety originated in the eighties. More specifically, Hauer (1986) first explored Empirical Bayesian estimation procedures that enhanced the accuracy of estimates. In his research, it is shown that the Empirical Bayes approach yields better estimates of the expected number of road accidents. Ever since, many applications used in some ways an Empirical Bayes approach [2]. Hauer and Persaud (1984) examined the performance of some identification procedures. Empirical Bayes methods were used to estimate proportions of correctly and falsely identified deviant road sections. Higle

[2] Examples referred to by Schlüter et al. (1997) and Van den Bossche et al. (2002)

and Witowski (1988) specified an upper limit value on the acceptable' underlying accident rate and identified a site i as being hazardous when the probability that the underlying accident rate exceeded the limit value by a predetermined tolerance level. Davies (1990) proposed a procedure for ranking a set of entities by considering a ratio between the underlying accident rate at each entity (target site) and the pooled underlying accident rates of the remaining entities (reference sites). Belanger (1994) applied Empirical Bayes methods to estimate the safety of four-legged unsignalized intersections. The results were used to identify black spot locations. Hauer (1996) reviewed the development of procedures to identify hazardous locations in general. Vogelesang (1996) gives a very comprehensive overview of Empirical Bayes methods in road safety research.

Another form of the Bayesian approach, called the hierarchical Bayesian model, has also been proposed in literature. These methods can handle the uncertainty and the great variability of accident data and produce a probabilistic ranking of the accident locations. However, the use of these models in traffic safety is less widespread [3]. Christiansen et al. (1992) developed a hierarchical Poisson regression model to estimate and rank accident sites, using a modified posterior accident rate estimate as a selection criterion. Schlüter et al. (1997) deal with the problem of selecting a subset of accident sites based on a probability assertion that the worst sites are selected first. They propose different criteria for site selection. To estimate accident frequencies, a hierarchical Bayesian Poisson model has been used. Davis and Yang (2001) combined hierarchical Bayes methods with an induced exposure model to identify intersections where the crash risk for a subgroup is relatively high. Tunaru (2002) proposed a hierarchical Bayesian approach for ranking accident sites based on a bivariate Poisson log-normal distribution.

The main advantage of these Bayesian statistical models is that they do not only take into account the available data to make estimates, they also allow to include prior information in the analysis. This prior information contains expert knowledge that is not based on the data that are available for the analysis. This will be explained in the following section.

3.5.2.2 Bayes Rule

In this study λ is the unknown parameter that represents the expected number of accidents for a location. For each location, this value needs to be estimated so that

[3] Examples referred to by Schlüter et al. (1997) and Brijs et al. (2003)

we can use these values to rank the locations.

Suppose the vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents the available information (i.e. the actual number of accidents) on the n locations we want to rank according to their expected accident rate. The n parameters that we are interested in, can then be represented by the vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)$. However, as explained in the previous section, the available data or empirical accident rate alone will not be sufficient to make reliable estimates for the future number of accidents at these locations. Therefore, we will include prior knowledge or assumptions on the values of $\boldsymbol{\lambda}$ in the prior distribution function $f(\boldsymbol{\lambda})$. Furthermore, as shown in (3.3), the distribution function of the data will depend on the values of $\boldsymbol{\lambda}$. This dependence can be expressed in a conditional distribution of probability $f(\boldsymbol{x} \mid \boldsymbol{\lambda})$. After taking into account this prior knowledge on the values of $\boldsymbol{\lambda}$ and the distribution function of the data, we can then use Bayes rule to estimate the values of $\boldsymbol{\lambda}$ (Spiegelhalter, 1999):

$$f(\boldsymbol{\lambda} \mid \boldsymbol{x}) = \frac{f(\boldsymbol{\lambda}) f(\boldsymbol{x} \mid \boldsymbol{\lambda})}{\int f(\boldsymbol{\lambda}) f(\boldsymbol{x} \mid \boldsymbol{\lambda}) d\boldsymbol{\lambda}} \approx f(\boldsymbol{\lambda}) f(\boldsymbol{x} \mid \boldsymbol{\lambda}) \tag{3.5}$$

In (3.5), $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$ is called the posterior or posteriori distribution for $\boldsymbol{\lambda}$. This distribution expresses the insecurity about $\boldsymbol{\lambda}$ after taking into account the available data. As explained above, $f(\boldsymbol{\lambda})$ is called the prior distribution function. Finally, $f(\boldsymbol{x} \mid \boldsymbol{\lambda})$ is called the likelihood. This is considered as the joint distribution of the data in function of the parameters λ_i. This distribution will integrate to 1. Consequently, in a simplified manner, Bayes Rules can be thought of in the following manner (3.6):

$$posterior = likelihood \times prior \tag{3.6}$$

From this joint posterior distribution $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$, the conditional posterior distributions for the λ_i's can be calculated for each individual location. This will be explained in the following sections.

3.5.2.3 Bayesian Hierarchical Models

In this research, we assume that each location i is characterized by a number of road accidents that are Poisson distributed with parameter λ_i , i.e. the expected number of accidents (see equation (3.3)). To estimate this value of λ_i, we use the empirical accident frequency that is available from the accident data. This information will be taken into account by the model in the form of the likelihood. Furthermore, we will include prior knowledge on the value of λ_i to estimate the expected accident rate for each location.

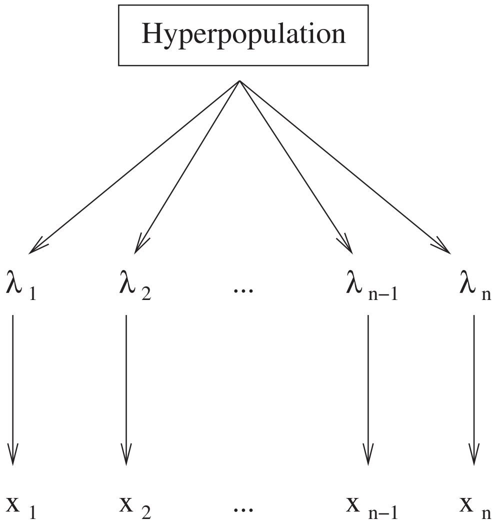

Figure 3.5: Diagram of the Hierarchical Bayesian Model

However, according to de Finetti (1974), it is generally unrealistic to assume that prior information can be readily and reliably elected for each individual site under investigation. As an alternative, Schlüter et al. (1997) proposed including expert prior information pertaining to characteristics of the grouping of accident sites under investigation. This information is both more readily available and more reliable. This feature can be implemented in the model by assuming exchangeability among the underlying accident rates. This implies that although each location has specific site characteristics, it is assumed that the accident rates on these locations arise from some common distribution, called the hyperdistribution.

Van de Bossche et al. (2002) explain that, based on Gelman et al. (1995), this setting can be schematically represented as in figure (3.5). Here, the value x_i (i=1,2,...,n) is the observed number of accidents at location i. This value depends on the value of the parameter λ_i, which represents the corresponding non-observable expected number of road accidents. This λ_i, in turn, originates from a hyperdistribution specified by some hyperparameters.

The authors further explain, that once the hierarchical structure is determined, a prior distribution should be chosen for the hyper population and its parameters. When there is no information available for the hyper parameters, a non-informative prior is preferred. In practice, this often implies assigning a Uniform or Normal distribution to the hyperparameter, in order to give each value of the hyperparameter the same probability. Specifying such a prior distribution for the hyper parameters, to express the uncertainty about their true value, is called the full Bayesian approach. A second frequently used approach, the Empirical Bayes approach, obtains point estimates for the parameters of the hyper prior from the data. In this case the priors are called informative priors.

For practical reasons, as in Brijs et al. (2003), in this research, we will make use of a Bayesian hierarchical model with informative priors because, especially for small counts, diffuse priors can have serious effects on the convergence properties of the chain (see following section). More specifically, these hyperparameters are estimated from the data by empirical Bayes.

3.5.3 MCMC Techniques and Gibbs Sampling

3.5.3.1 Markov Chain Monte Carlo (MCMC) Techniques

The main interest in this research is in the posterior probability for road accidents at each location. In the previous section, the joint posterior distribution of these probabilities is given by equation (3.5). Here, $\boldsymbol{\lambda}$ is the vector of the unknown expected number of accidents and $\mathbf{x}$ is the vector of the observed number of road accidents (data). Since we are interested in ranking the locations based on their expected number of accidents, information is required on the single values of the λ's for each location, say λ_k. This is achieved by integrating the joint posterior distribution over all other parameters λ_i (i = 1, 2, ..., k-1, k+1,..., n) (Spiegelhalter, 1999):

$$f(\lambda_k \mid \boldsymbol{x}) = \int f(\boldsymbol{\lambda} \mid \boldsymbol{x}) d\lambda_1 d\lambda_2 ... d\lambda_{k-1} d\lambda_{k+1} ... d\lambda_n \tag{3.7}$$

One way to solve these marginal posterior probabilities is to work out the integrals analytically, for which the calculations may become very cumbersome (Van den Bossche et al., 2002). Another approach is to sample from the posterior distribution. This is the basic principle of Monte Carlo simulation. Numerical calculations are carried out using simulation. Instead of analytically calculating exact or approximate estimates, the Markov Chain Monte Carlo (MCMC) technique generates a stream of simulated values for the quantities of interest (Spiegelhalter (1996), Congdon (2003)).

In general, samples should be drawn from the joint posterior distribution $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$. MCMC simulation is used when it is not possible (or not computationally efficient) to sample directly from the joint posterior distribution. Instead, we sample iteratively in such a way that each step of the process we expect to draw from a distribution that becomes closer and closer to the distribution function. More specifically, MCMC simulation is a general method based on drawing values of λ from approximate distributions and then correcting those draws to better approximate $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$. The samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn; hence the draws form a Markov chain. (As defined in probability theory, a Markov chain is a sequence of random variables θ^1, θ^2, ..., for which, for any t, the distribution of θ^t given all previous θ's depends only on the most recent value, θ^{t-1}). For a wide class of problems, including posterior distributions for many hierarchical models, this approach appears to be the easiest way to get reliable results when used carefully. The key is to create a Markov Process whose stationary distribution is the specified $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$ and run the simulation long enough that the distribution of the current draws is close enough to this stationary distribution. This is called 'convergence' (Gelman et al., 1995).

3.5.3.2 Gibbs Sampling

A particular Markov chain algorithm that has been found to be useful in many multidimensional problems is the Gibbs sampler, also called alternating conditional sampling, which is defined in terms of subvectors of λ (Gelman et al., 1995).

If the vector of parameters λ consists of n sub-components, then starting values $\lambda_1^{(0)}$, $\lambda_2^{(0)}$,..., $\lambda_n^{(0)}$ should be defined. The following sample scheme is then repeated thousands of times, to eventually obtain a sample from $f(\boldsymbol{\lambda} \mid \boldsymbol{x})$, the stationary distribution of the Markov chain:

Sample $\lambda_1^{(1)}$ from $f(\lambda_1 \mid \lambda_2^{(0)}, \lambda_3^{(0)}, ..., \lambda_n^{(0)}, \mathbf{x})$;
Sample $\lambda_2^{(1)}$ from $f(\lambda_2 \mid \lambda_1^{(1)}, \lambda_3^{(0)}, ..., \lambda_n^{(0)}, \mathbf{x})$;
...
Sample $\lambda_n^{(1)}$ from $f(\lambda_n \mid \lambda_1^{(1)}, \lambda_2^{(1)}, ..., \lambda_{n-1}^{(1)}, \mathbf{x})$.

After a large number of iterations, the distribution of the value of λ for each accident site will have been sampled. Posterior summary statistics may be calculated to infer the newly obtained distributions.

One particular application of the Gibbs sampler is the ability to make inferences on arbitrary functions of unknown parameters (Spiegelhalter, 1996). An example of this feature is the computation of the rank probability for each location. The result is a sample from the posterior distribution of the ranks (Van den Bossche et al., 2002).

More specifically, for every iteration, the Gibbs sampler provides an estimate of the rank for a given location i by scanning the estimated number of accidents of all other accident locations at that iteration. The number of locations with a sampled value for λ that is equal or lower than λ_i corresponds to the rank of location$i$.

3.5.4 Empirical Study

3.5.4.1 Multivariate Hierarchical Bayes Approach

In this dissertation, we followed the approach of Brijs et al. (2003), who proposed a multivariate hierarchical Bayes approach with informative priors for ranking accidents sites. More specifically, suppose the data consist of n different sites. The number of accidents for the i-th site for a fixed time period (here this time period equals 3 years, i.e. 1997-1999) is denoted as X_i. We assume that the number of accidents for this site follows a Poisson distribution with parameter λ_i. For each site, we have also the triplets (D_i, S_i, L_i) that respectively correspond to the number of deadly injured persons or fatalities, the number of seriously injured persons and the number of lightly injured persons for the given time period for each site. We assume that jointly and conditional on the number of accidents X_i, they follow a 3-variate Poisson distribution.

Furthermore, in this model we allow for pairwise covariances for each pair of variables instead of the usual model that assumes the same covariance term for all the pairs of variables. In other words, the model allows for different correlations between the variables D_i, S_i and L_i, which is a more realistic assumption in the context of traffic accident injuries.

This results is the following model:

$$\begin{array}{rcl} X_i & \sim & Poisson(\lambda_i) \\ (D_i, S_i, L_i) \mid X_i = x_i & \sim & 3 - Poisson(\mu_{1i}xi, \mu_{2i}x_i, \mu_{3i}x_i, \lambda_{12}x_i, \lambda_{13}x_i, \lambda_{23}x_i) \end{array}$$

Hence, μ_i reflects the rate for fatalities, serious injuries and light injuries per accident for the site i, while λ_{ij} are the covariance parameters for each pair of variables.

Brijs et al. (2003) note that empirical evidence supports the assumption that there is a positive correlation between the three variables D_i, S_i and L_i. They explain that this is natural since it reflects the severity of the accidents on location i. So, instead of assuming independence between the three variables, by imposing three independent Poisson distributions, a model is proposed that takes into account those correlations between the variables, and hence it can model the interdependencies in a more realistic way.

In order to combine all data into a single number that will be used for ranking the sites, a weighing function can be used that measures the expected score of an accident according to the number of fatalities, seriously and lightly injured persons. Based on these expected scores, the posterior density for the rank of each site can be derived. The parameters of this model are estimated via Bayesian estimation facilitated by Markov Chain Monte Carlo (MCMC) methods (a more detailed description of this technique can also be found in Brijs et al. (2003)).

This approach allows to investigate the effects of using the expected number of accidents, estimated from a hierarchical Bayesian model, instead of the historic count data to rank and select the most dangerous accidents sites. It also allows to use several cost functions based on different weighing value combinations.

3.5.4.2 Results for the 1,014 Currently Most Dangerous Accident Locations

Table 3.11 shows the results of the comparative analysis between the location rankings based on count values and the location rankings based on Bayesian estimates for the 1,014 accident locations that are currently considered as most dangerous.

More specifically, table 3.11 gives, for different subsets of the data set, the percentage deviation values which refer to the number of accident locations that differ between the ranking based on the count values and the ranking based on the Bayesian estimates (see equation 3.2). The values on the diagonals of this table represent the effect of using Bayesian estimation values instead of historic count data while using the same weight value combination to select different subsets of the 1,014 accident locations.

These results show that using Bayesian estimation techniques instead of the historic count data while using the same weight value combination can lead to the selection of 3.9% to 10.6% different accident locations. These results are shown on the diagonals of the table. When selecting the 800 most dangerous accident locations using the 1_ 3_ 5 weighing values this leads to the selection of 8.5% or 68 different accident locations. Translated into costs, this means that theoretically 42.5 million EURO of the 500 million EURO investment budget for redesign is allocated partly due to the random variation in accident counts.

Additionally, the results of table 3.11 enable to estimate the effect of a change in

Table 3.11: Comparative analysis between counts and Bayes estimates for the 1,014 currently most dangerous accident locations

D for top 15%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	7.3%	29.4%	27.4%	50.3%
	1_ 1_ 10	31.4%	5.9%	24.2%	45.1%
	1_ 3_ 5	24.8%	22.9%	5.9%	29.4%
	1_ 10_ 10	45.1%	40.8%	20.3%	7.8%
D for top 40%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	5.9%	27.8%	25.9%	48.0%
	1_ 1_ 10	27.1%	5.9%	24.9%	44.6%
	1_ 3_ 5	19.2%	23.9%	10.6%	31.0%
	1_ 10_ 10	39.9%	39.6%	17.5%	10.6%
D for top 70%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	5.5%	18.0%	23.8%	34.8%
	1_ 1_ 10	16.8%	5.1%	19.1%	28.2%
	1_ 3_ 5	15.8%	15.2%	10.3%	19.6%
	1_ 10_ 10	31.3%	15.5%	12.8%	4.1%
D for top 800		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	5.12%	13.0%	23.6%	25.7%
	1_ 1_ 10	12.7%	5.4%	15.5%	23.7%
	1_ 3_ 5	13.2%	11.4%	8.5%	15.6%
	1_ 10_ 10	23.1%	21.5%	10.7%	3.9%

the parameter weights combined with the use of Bayesian estimation values instead of count data. For example, from table 3.1 it is shown that when selecting the 15% most dangerous accident locations a change in the parameter values from 1_ 1_ 1 to 1_ 10_ 10 can lead to a selection of 43.8% different accident locations. When also using Bayesian estimation, table 3.11 shows that this can even result in the selection of 45.1% different accident locations when comparing the rankings of the 1_ 1_ 1 Bayesian estimations with the ranking of the 1_ 10_ 10 count values, and even 50.3%

Table 3.12: Total number of fatal, serious and light injuries

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	245	2,364	13,173	15,782
1_ 1_ 10	374	2,326	12,708	15,408
1_ 3_ 5	354	2,650	12,015	15,019
1_ 10_ 10	355	2,809	11,127	14,291

different accident locations when comparing the rankings of the 1_ 1_ 1 count data with the ranking of the 1_ 10_ 10 Bayesian estimations. Note that these percentage deviation values will be higher that the percentage deviation values reflected on the diagonals of the table, which naturally can be explained by the effect of changing the weight value combinations. Conform to the results of table 3.1 the percentage deviation values will also be smaller when more accident locations are selected.

Finally, as explained before (see table 3.2), using count data and 1_ 3_ 5 weights, selecting and tackling the 800 most dangerous accident locations relates to 329 deadly injured victims (DI), 2,571 seriously injured persons (SI) and 12,496 lightly injured persons (LI). When using Bayesian estimation values instead of count data, depending on the chosen alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 245 to 374), seriously injured persons (ranging from 2,364 to 2,809) and lightly injured persons (ranging from 11,127 to 13,173). The total number of injured persons will then vary from 14,291 to 15,782 (see table 3.12. As explained before, this choice between avoiding as much victims as possible or prioritizing more serious injury accidents represent different attitudes from the government towards the traffic safety problem.

3.5.4.3 Results for all 23,184 Accident Locations

In table 3.13 the results of the comparative analysis between the location rankings based on count values and the location rankings based on Bayesian estimates for all 23,184 accident locations are presented.

Similar to the results presented in table 3.11, the values on the diagonals of this table represent the effect of using Bayesian estimation values instead of historic count data while using the same weight value combination to select different subsets of the 23,184 accident locations.

Table 3.13: Comparative analysis between counts and Bayes estimates for all accident locations

D for top 800		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	4.6%	15.7%	15.5%	36.3%
	1_ 1_ 10	34.7%	18.2%	27.0%	37.5%
	1_ 3_ 5	22.7%	18.3%	8.12%	17.3%
	1_ 10_ 10	46.0%	41.2%	31.0%	8.6%
D for top 15%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	5.1%	17.1%	17.4%	37.0%
	1_ 1_ 10	25.9%	9.1%	21.6%	35.6%
	1_ 3_ 5	23.5%	18.8%	6.8%	16.4%
	1_ 10_ 10	39.1%	34.1%	22.7%	3.4%
D for top 40%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	1.7%	5.9%	19.5%	39.4%
	1_ 1_ 10	5.9%	3.7%	15.6%	35.3%
	1_ 3_ 5	19.9%	15.8%	0.0%	19.5%
	1_ 10_ 10	39.4%	35.3%	19.9%	0.0%
D for top 70%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	10.2%	10.4%	10.1%	10.2%
	1_ 1_ 10	10.2%	9.3%	9.1%	9.2%
	1_ 3_ 5	10.3%	9.2%	0.1%	0.1%
	1_ 10_ 10	10.3%	9.2%	0.1%	0.1%

These results show that this ranking procedure can lead to the selection of approximately 0.0% different accident locations for the 1_ 3_ 5 and 1_ 10_ 10 weighing values when selecting the top 40% most dangerous accident locations, up to 18.2% different accident locations for the 1_ 1_ 10 weighing value combination when selecting the 800 most dangerous accident locations. These results show that selecting the 800 most dangerous accident locations based Bayesian estimation values, instead of historic count data, causes 18% or 90 million EURO of the investment budget for

Table 3.14: Total number of fatal, serious and light injuries

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	217	2,115	13,589	15,921
1_ 1_ 10	435	2,081	12,914	15,430
1_ 3_ 5	325	2,512	12,771	15,608
1_ 10_ 10	372	3,005	10,670	14,047

redesigning these accident locations to be differently selected.

Furthermore, results of table 3.13 show the effects of change in the parameter weights combined with the use of Bayes estimation instead of count data. The maximum percentage deviation value of this table shows that this ranking procedure can lead to the selection of 46.0% different accident locations when comparing the 1_ 1_ 1 Bayesian ranking with the 1_ 10_ 10 counts ranking for the 800 most dangerous accident locations. This result is slightly higher than the percentage deviation value of table 3 for the same subset and the same weighing value combinations. Note that again these percentage deviation values will be smaller when more accident locations are considered in the subset.

Finally, as shown in table 3.4, using count data and 1_ 3_ 5 weights, selecting and tackling the 800 most dangerous accident locations relates to 401 deadly injured victims (DI), 2,636 seriously injured persons (SI) and 12,141 lightly injured persons (LI). Analogously with table 3.12, table 3.14 shows that using Bayesian estimation values instead of count data and depending on the chosen alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 217 to 372), seriously injured persons (ranging from 2,081 to 3,005) and lightly injured persons (ranging from 10,670 to 13,589). The total number of injured persons will vary from 14,047 to 15,921.

3.5.4.4 Results for the 5,326 Accident Locations with Minimum 3 Accidents

Table 3.15 presents the results of a comparative analysis between ranking based on historic count data on the one hand and Bayesian estimation on the other hand for the accident locations where at least 3 accidents occurred in the last 3 years.

The results on the diagonals of table 3.15 show that using Bayes estimation instead of historic count data to rank and select the accident locations can lead to a maximum

Table 3.15: Comparative analysis between counts and Bayes estimates for all accident locations with minimum 3 accidents

D for top 15% (800)		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	4.1%	22.1%	22.1%	43.7%
	1_ 1_ 10	27.0%	5.1%	22.0%	38.6%
	1_ 3_ 5	18.4%	18.2%	4.5%	25.9%
	1_ 10_ 10	39.2%	34.0%	20.2%	5.2%
D for top 40%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	5.7%	11.5%	19.3%	30.2%
	1_ 1_ 10	9.7%	4.6%	15.1%	26.9%
	1_ 3_ 5	17.0%	14.3%	5.4%	15.1%
	1_ 10_ 10	27.1%	25.7%	12.2%	4.9%
D for top 70%		Bayes			
		1_ 1_ 1	1_ 1_ 10	1_ 3_ 5	1_ 10_ 10
Counts	1_ 1_ 1	0.3%	1.8%	12.5%	17.8%
	1_ 1_ 10	1.9%	1.5%	11.9%	17.1%
	1_ 3_ 5	11.1%	10.2%	2.9%	7.9%
	1_ 10_ 10	17.9%	17.0%	5.7%	2.5%

percentage deviation value of 5.7% for the 40% most dangerous accident locations using the 1_ 1_ 1 weighing values. Note that compared to the results of table 3.11 and table 3.13 this maximum value is relatively low. However, this still means that almost 6% of the accident locations that are considered to belong to the top 40% most dangerous accident locations are not considered as dangerous when using Bayesian estimation and vice versa.

Furthermore, the results of table 3.15 show that a change in the parameter weights combined with the use of Bayes estimation can lead to even higher percentage deviation values. For example, from table 3.5 it is shown that when selecting the 40% most dangerous accident locations a change in the parameter values from 1_ 1_ 1 to 1_ 10_ 10 can lead to a selection of 26.1% different accident locations. When also using Bayes estimation, table 3.15 shows that this can result in the selection of 27.1% different accident locations when comparing the rankings of the 1_ 1_ 1 Bayesian estimations

Table 3.16: Total number of fatal, serious and light injuries

Top 800	DI	SI	LI	TOTAL
1_ 1_ 1	228	2,159	13,530	15,917
1_ 1_ 10	444	2,109	12,529	15,082
1_ 3_ 5	332	2,616	12,301	15,249
1_ 10_ 10	344	2,973	10,003	13,320

with the ranking of the 1_ 10_ 10 count values, and even 30.2% different accident locations when comparing the rankings of the 1_ 1_ 1 count data with the ranking of the 1_ 10_ 10 Bayesian estimations. Analogously to the previous results, the percentage deviation values will decrease when more accident locations are selected in the subset.

Finally, as explained in table 3.6, using count data and 1_ 3_ 5 weights, selecting and tackling the 800 most dangerous accident sites covers 329 deadly injured victims (DI), 2,571 seriously injured persons (SI) and 2,496 lightly injured persons (LI). When using Bayesian estimation values instead of count data, depending on the chosen alternative weighing combination, the selection of the 800 most dangerous sites will cover a different number of respectively fatally injured persons (ranging from 228 to 444), seriously injured persons (ranging from 2,109 to 2,973) and lightly injured persons (ranging from 10,003 to 13,530) (see table 3.14). The total number of injured persons will vary from 13,320 to 15,917.

3.5.5 Conclusions

In this research, we used a multivariate hierarchical Bayes approach for ranking and selecting accidents sites taking into account the number of accidents, the number of fatalities, and the number of lightly and severely injured casualties for a given time period for each site. This approach takes into account the random variation in accident counts by using a 3-variate Poisson distribution that allows for covariance between the variables. In order to combine all data into a single number that will be used for ranking the sites, a weighing value combination is used to measure the expected score of an accident according to the number of fatalities, heavy and light injured casualties. Based on these expected scores, the posterior density for the rank of each site can be derived.

Results showed that depending on the chosen weighing value combination, selecting the 800 most dangerous sites using Bayesian estimation values instead of the historic count data can lead to the selection of up to 18.2% different accident locations. These results illustrate that, in comparison with the use of accident count data, the use of Bayesian estimation techniques can cause different accident sites to be considered as dangerous. This will have a great impact on the ranking and selection of dangerous accident locations, and accordingly on the allocation of the investment budget for the most dangerous accident sites.

3.6 Probability Plots

3.6.1 Bayesian Ranking Plots

In this section, we elaborate on the technique of Bayesian estimation that was discussed in the previous section. More specifically, we propose a method to generate Bayesian ranking plots in order to visualize the probability that a location will be ranked as dangerous, based on estimates from a hierarchical Bayes model. We will explore the possibility of using these probability plots as a graphical instrument to select dangerous locations.

In particular, we will derive the estimated probability for each site i of being one of the r most dangerous sites (with n = the total number of locations). This implies that the expected score of location i (using the weighing value combination) is among the r highest and hence its rank R is larger than (n - r) (since in this ranking procedure the larger the value of R, the worse the site). Then, the estimated probability $P_r(i)$ is calculated as (3.8):

$$P_r(i) = \frac{\sum_{j=1}^{N} I(R_j^{(i)} > n - r)}{N} \tag{3.8}$$

where I is the indicator function returning a value of 1 in case that the argument is true and a value of 0 in case that the argument is false. N is the number of MCMC iterations. These probabilities allow for a heuristic rule for selecting dangerous sites. More specifically, if all sites would have the same characteristics, we expect that for all the sites the required probabilities of belonging to the worst sites will be exactly the same as any differences will be merely random perturbations. Accordingly, we expect that this probability will be equal to r/n for each site. Locations with a probability above this limit reveal a deviation from the argument about equal sites. However, note that theoretically, due to random perturbations some probabilities will be larger even in the case of equal sites.

As this technique allows to estimate the probability that a location will be ranked as belonging to the r most dangerous sites, this implies that for this research, we can use this technique to estimate the probability that a location will be ranked as one of 800 most dangerous sites. These probabilities can then be visualized using probability plots.

Furthermore, we can calculate 'confidence' intervals for these probabilities, by repeating the above procedure for a number of times and taking into account the minimum and maximum generated probability for each site. Indeed, in order to take into account as much as possible the instability and variability that characterize the accident counts, the model does not generate exactly one ranking order for each accident location. Instead, for each location, the model produces a series of expected ranking orders (one for each iteration), given that the number of accidents at each location will fluctuate around a mean value that is typical for this location.

In practice, this corresponds with splitting up the total number of MCMC iterations (N) in a number of batches and calculate the estimated probability for each site after each batch. This will allow generating Bayesian confidence intervals for each site. By considering the lower limit of these intervals (the smallest generated value for the probability of belonging to the worst sites), this will reveal sites with a probability above the limit in a more rigorous way, thus reducing the effect of random perturbations. In other words, by selecting these sites, we select the locations that, for every iteration of our model, have a higher probability than expected under random conditions to belong to the 800 most dangerous accident locations.

3.6.2 Empirical Results

3.6.2.1 Results for all 23,184 Accident Locations

Based on the estimated score for each accident location, obtained from the hierarchical Bayes model, we are able to estimate the probability for each accident location to belong to the 'r' most dangerous locations.

For instance, the curve in figure (3.6) shows for each of the 23,184 road locations (the X-axis) the estimated probability of belonging to the 800 most dangerous accident locations (the Y-axis), ordered by decreasing probability.

Additionally, the horizontal line in figure (3.6) shows that if each location would be equally dangerous, accidents would occur randomly on the different locations. In

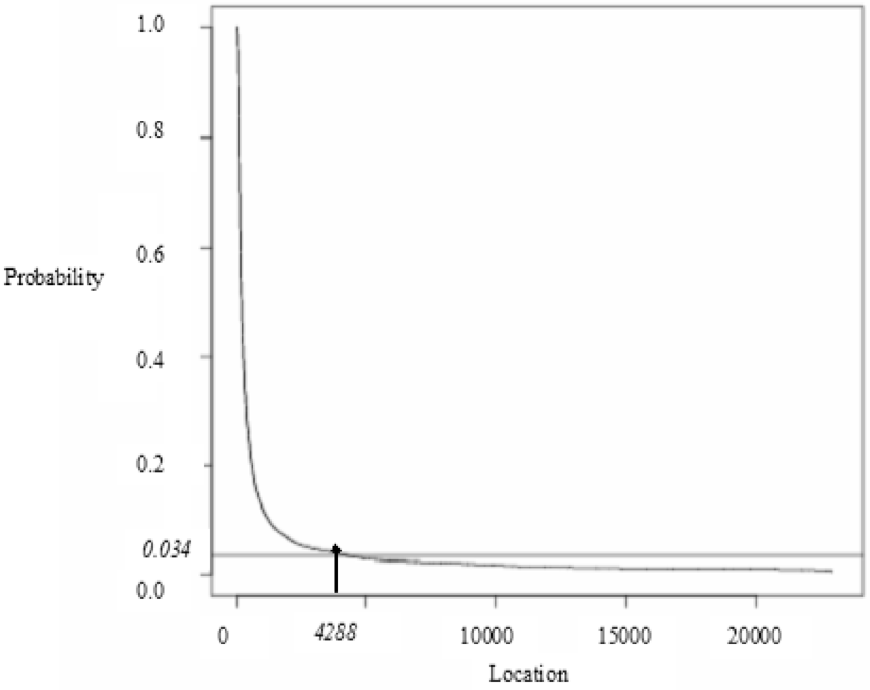

Figure 3.6: Bayesian ranking plot: Probability of belonging to the 800 most dangerous of all accident locations

that case, the probability that a location belongs to the 800 most dangerous accident locations would be equal for all accident locations, namely 800/ 23,184 = 0.034. It can be seen that this value is relatively small due to the large number of accident locations included in this data set.

However, the curved line in figure 2 shows that the probability of belonging to the 800 most dangerous accident locations is not at all equal for the 23,184 locations included in the data set. More specifically, 4,288 locations have a probability that is larger than 0.034. These locations can be identified in figure (3.6) as those locations for which the curve is above the horizontal cutoff line. This indicates that these accident locations have a higher probability than expected under random conditions to qualify as one of the 800 most dangerous accident locations.

When comparing the 800 locations with the highest estimated probabilities based on the results from figure (3.6) with the 800 locations that are currently considered as the most dangerous, we found a percentage deviation value of 13.7%. This corresponds with 110 accident locations that are differently selected when targeting the 800 most dangerous accident sites. Translated into costs, this means that theoretically 68.5 million EURO of the 500 million EURO investment budget for redesigning these 800 most dangerous accident locations would be differently allocated. Closer investigation of these different accident locations shows that these sites, according to the currently used ranking criterion, are ranked between the 500th and 800th position. In other words, the locations that are currently considered as belonging to the 500 most dangerous also have the highest estimated probabilities of belonging to the 800 most dangerous sites.

Figure (3.7) shows the results of generating 'confidence intervals' for each of the 23,184 locations. More specifically, the vertical lines in this picture represent for each accident site the minimum and maximum estimated probability to belong to the 800

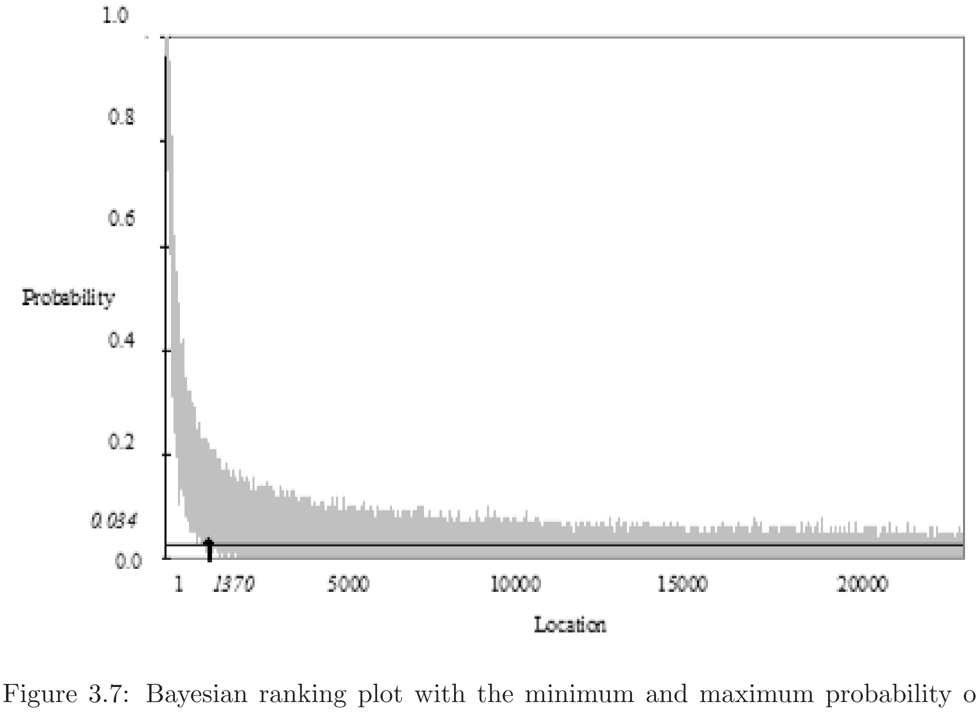

Figure 3.7: Bayesian ranking plot with the minimum and maximum probability of belonging to the 800 most dangerous of all accident locations

most dangerous locations out of the 50 MCMC batches that were included in this analysis. Note that the mean estimated probability for each accident site from the different iterations will equal the estimated probability depicted in figure (3.6).

These results show that selecting the locations with the lower limit of their confidence interval (the minimum estimated probability) above the limit of 0.034 results in 1,370 accident locations. This indicates that these accident locations have a higher probability than expected under random conditions to qualify as one of the 800 most dangerous accident locations.

Furthermore, comparing these results with the results of figure (3.6) shows that using confidence intervals based on the estimated probabilities reduces the selection of the candidate 800 most dangerous locations from 4,288 accident sites to 1,370. Consequently, the use of confidence intervals results in a more conservative estimate of the most dangerous accident locations, which for policy makers enhances the certainty that resources are allocated to the right accident sites.

3.6.2.2 Results for the 5,326 Accident Locations with Minimum 3 Accidents

Analogously to figure (3.6), figure (3.8) shows for each of the 5,326 locations where minimum 3 accidents occurred between 1997 and 1999 (X-axis) the probability that it belongs to the 800 most dangerous accident locations (Y-axis).

More specifically, these probabilities are depicted in a curved line, while the horizontal line in figure (3.8) represents this probability under the assumption that all sites would be equally dangerous. This assumption results in a probability value of 0.15 (800/ 5,326). Obviously, this value is larger than the probability value calculated in figure (3.6), since now only 5,326 accident locations are included in the analysis. Comparing this value with the probabilities estimated from the hierarchical Bayes model, shows that 1 506 accident locations have a probability of belonging to the 800 most dangerous sites. This is higher than what would be expected in case of equally dangerous sites.

Additionally, selecting the 800 accident locations with the highest probabilities and comparing these sites with the currently 800 most dangerous locations results in a percentage deviation value of 12%. This means that 96 locations that are currently considered to belong to the 800 most dangerous accident sites will not be selected in this group based on their estimated probabilities. Similar to the results of figure (3.6), the accident locations that are selected differently are ranked between the 500th and

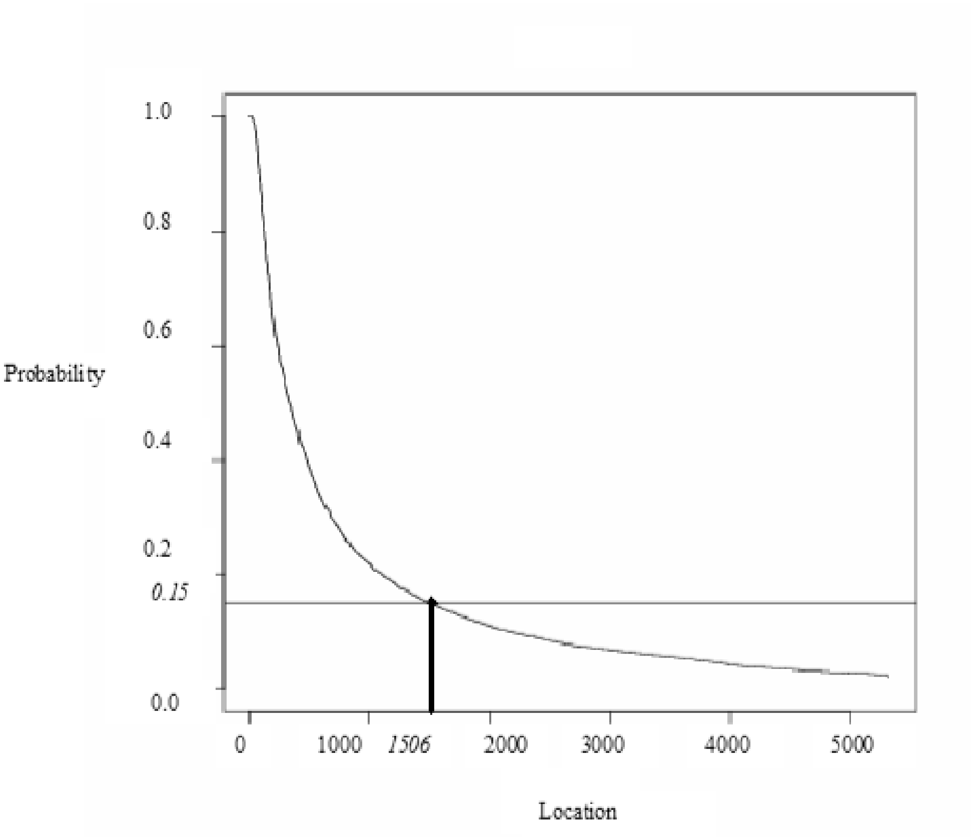

Figure 3.8: Bayesian ranking plot: Probability of belonging to the 800 most dangerous of the locations with minimum 3 accidents

800th position according to the currently used ranking criterion. When translating these results into costs, this indicates that theoretically 60 million EURO of the 500 million EURO investment budget for redesigning these 800 most dangerous accident locations would be differently allocated.

In figure (3.9), for each accident location the minimum and maximum estimated probability of the different iterations is shown resulting in confidence intervals'. Analogously to figure (3.7), the mean estimated probability for each site will equal the estimated probability depicted in figure (3.8).

Results of figure (3.9) show that for 863 accident locations the minimum estimated probability value of belonging to the 800 most dangerous accident locations exceeds the limit of 0.15. In other words, these accident sites have a higher probability than

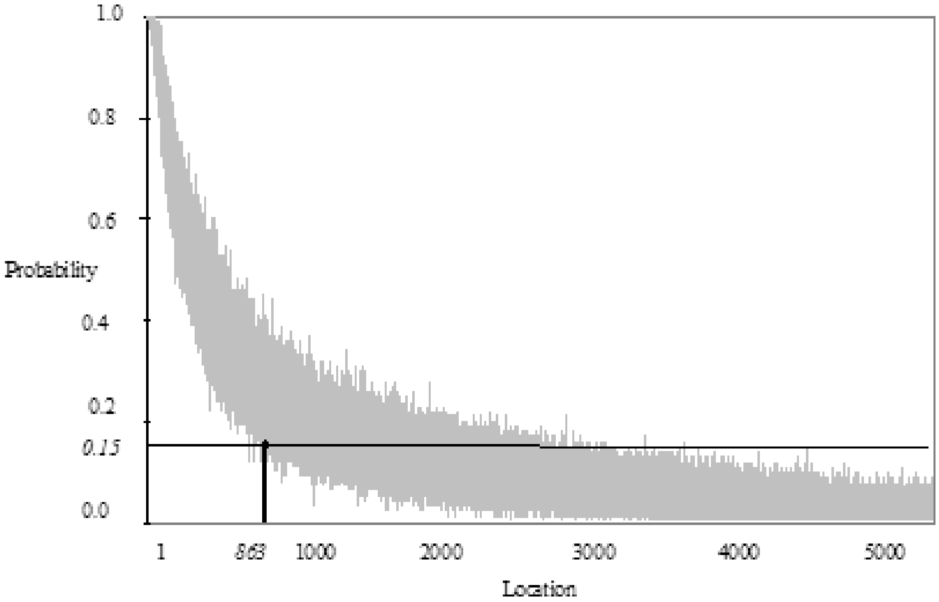

Figure 3.9: Bayesian ranking plot with the minimum and maximum probability of belonging to the 800 most dangerous of the locations with minimum 3 accidents

expected under random conditions to qualify as one of 800 most dangerous accident locations.

Furthermore, results of figure (3.8) and figure (3.9) show that using the lower limit of the 'confidence intervals' to select the accident locations with an estimated probability of belonging to the 800 most dangerous accident locations narrows down the number of sites from 1,506 to 863. Again, these results indicate that 'confidence intervals' facilitate a more rigorous estimate of the most dangerous accident locations.

Finally, note that, in contrast with the previous section, the technique of Bayesian ranking plots is not applied to the 1,014 accident locations that are currently considered as dangerous. This is motivated by the fact that estimates for the probability of belonging to the 800 most dangerous accident locations are not very interesting when dealing with just 1,014 locations. More specifically, when selecting 800 locations out of 1,014 sites, the probability of being selected (800/1,014) will be intrinsically high (almost 80%).

3.6.3 Conclusions

Based on estimates from a hierarchical Bayes model, Bayesian ranking plots can be used to visualize the estimated probability that a location will be ranked as belonging to the r most dangerous locations. These probability plots can provide policy makers with a scientific instrument with intuitive appeal to select dangerous road locations on a statistically sound basis. In particular, this implies that we can use this technique to estimate the probability that a location will be ranked as one of the 800 most dangerous sites. Additionally, in view of the instability and variability that characterize the accident counts, this model is able to produce a series of ranking orders that can be expected taking into account that the number of accidents at each location will fluctuate around a mean value that is typical for this location.

3.7 Combining the Different Impacts: Case Study

3.7.1 Valuation of Casualties

In this study, a sensitivity analysis is performed to investigate how big the impact would be on the ranking and selection of dangerous accident locations in Flanders when we combine the three different selection criteria discussed above: (1) using alternative weighing value combinations, (2) weighing the severity of the accident instead of all the injured occupants and (3) using Bayesian estimation values instead of historic count data. More specifically, in this case study, we will only take into account the most serious injury per accident and use a valuation of casualties based on direct costs, indirect costs and validation for human suffering to rank the accident locations.

In particular, the weighing values used in this section are based on accident costs which are often used in cost-benefit analyses to value the impact of road safety measures in Norway (Elvik, 2004). These accident costs were estimated by Elvik in 1993 and are the sum of five main items: medical costs, loss of output, costs of property damage, administrative costs, economic costs and economic valuation of lost quality of life. This sum results in a total cost of minimum respectively 16,600,000, 3,780,000 and 500,000 per respectively fatally injured, seriously injured and lightly injured person (1995 prices, Norwegian kroner). Converting these costs into cost ratios between the different injury types results in the weighing value combination 1_7_33.

More specifically, as explained in section 3.4.1, in practice this means that the points per accident that are summed up in order to calculate the priority value of the locations can vary between 1 (only light injuries), 7 (at most serious injuries) and 33 (at least one deadly injured casualty). These values represent the difference in costs that can be avoided by preventing these injuries from happening. Therefore, we will use these weighing values in our analysis as an alternative for the 1_ 3_ 5 weighing values to calculate the priority score for each accident location.

Using these weighing values, we will generate probability plots, based on estimates from a hierarchical Bayes model, in order to visualize the estimated probability that a location will be ranked as dangerous.

For this case study, we only select the sites where at least 3 accidents occurred between 1997 and 1999. This results in 5,326 accident locations that will be analyzed in this research.

3.7.2 Empirical Results

Analogously to figure (3.8), figure (3.10) shows for each of the 5,326 locations where minimum 3 accidents occurred (X-as) the probability of belonging to the 800 most dangerous accident locations, ordered by decreasing probability. Again, if all sites were equally dangerous and accidents would occur randomly on the different locations, the probability that a location belongs to the 800 most dangerous accident locations would be equal for all accident locations, namely 800/5,326 = 0.15. This is represented by the horizontal line in figure (3.10).

However, from the curved line in figure (3.10), it can be seen that the probability of belonging to the 800 most dangerous accident locations is not at all equal for the 5,326 locations with minimum 3 accidents. More specifically, 1,431 locations have a probability that is larger than 0.15. These locations can be identified in figure (3.10) as those locations for which the curve is above the horizontal cut-off line. Remember, that this indicates that these accident locations have a higher probability than expected under random conditions to qualify as one of the 800 most dangerous accident locations.

When comparing the 800 accident locations that are currently considered as dangerous (using 1_ 3_ 5 weights for all the casualties, based on the empirical accident rate) with these 1,431 locations, it turns out that only 653 of the 800 current dangerous accident locations have a probability that is larger than 0.15. In other words, 147 accident locations are currently considered as belonging to the 800 most dangerous locations while according to the Bayesian ranking technique, (taking into account

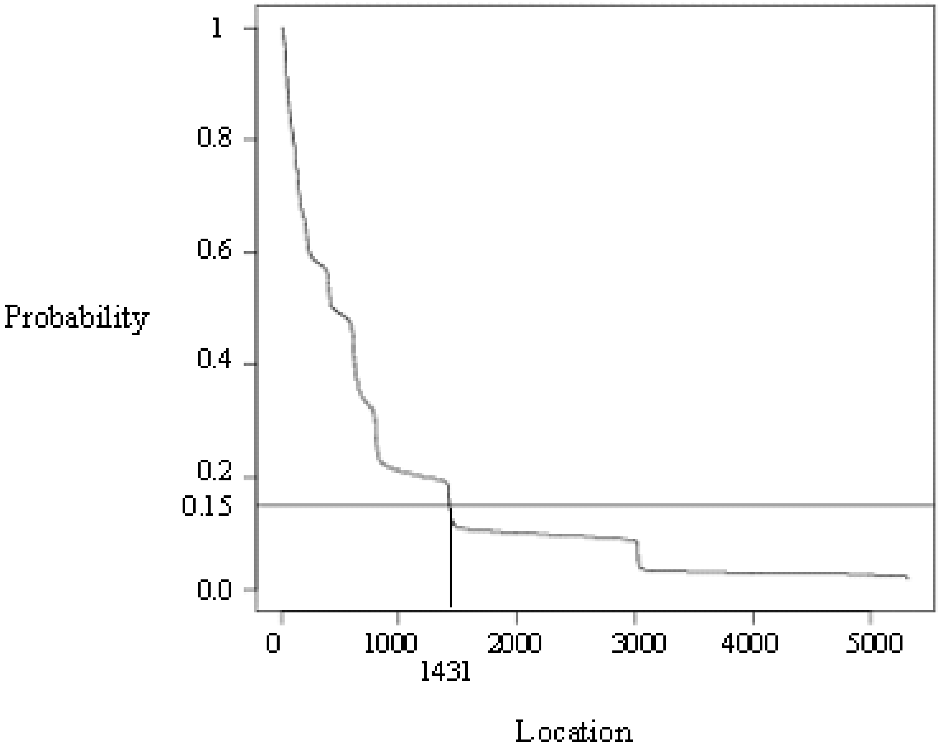

Figure 3.10: Bayesian ranking plot case study: Probability of belonging to the 800 most dangerous accident sites

the most serious injury per accident and using a valuation of casualties based on direct costs, indirect costs and validation for human suffering) the probability for these locations is lower than expected under random conditions.

Furthermore, selecting the 800 locations with the highest estimated probabilities based on the results from figure (3.10) and comparing these sites with the 800 locations that are identified according to the Flemish selection procedure, results in a percentage deviation value of 40.6%. This corresponds with 325 accident locations that are differently selected when targeting the 800 most dangerous accident sites. Translated into costs, this means that theoretically 205 million EURO of the 500 million EURO investment budget for redesigning these 800 most dangerous accident locations would be differently allocated when using the ranking criteria proposed in this case study.

In figure (3.11), for each accident location the minimum and maximum estimated

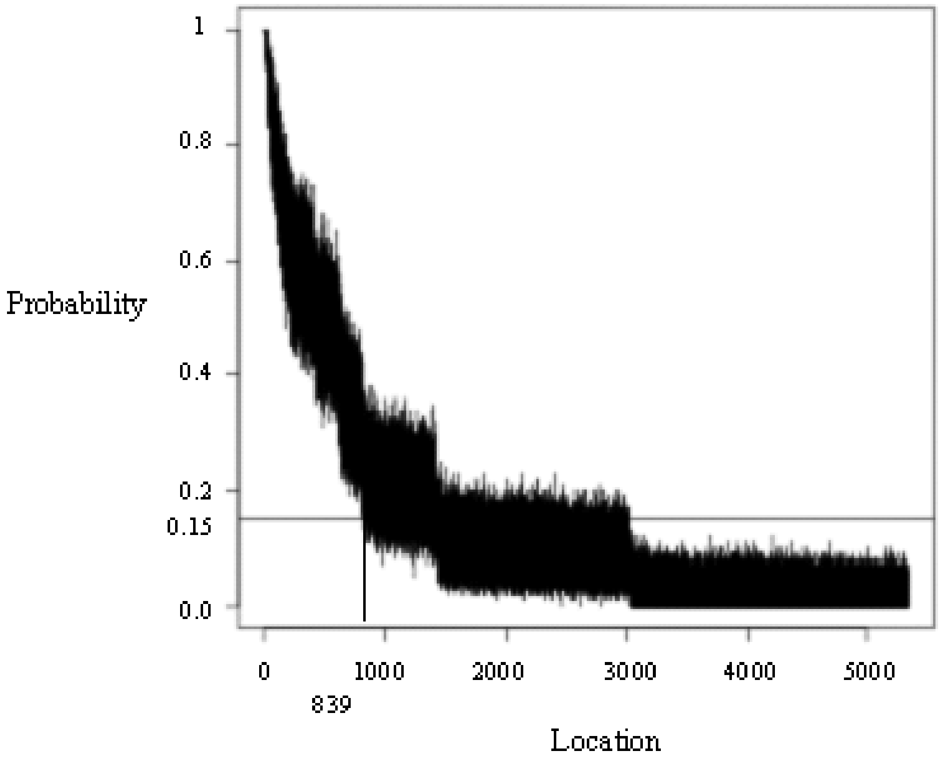

Figure 3.11: Bayesian ranking plot case study: Minimum and maximum probability of belonging to the 800 most dangerous accident sites

probability of belonging to the 800 most dangerous locations across the different batches of MCMC iterations is shown resulting in the vertical line in the picture. This way we create some sort of 'confidence interval' for each location. Note that the mean estimated probability for each accident site from the different iterations will equal the estimated probability depicted in figure (3.10).

These results show that for 839 accident locations the minimum estimated probability value of belonging to the 800 most dangerous accident locations exceeds the limit of 0.15. In other words, by incorporating as much variability as possible and accordingly selecting as strict as possible, 839 accident sites have a probability that is always higher than expected under random conditions to qualify as one of 800 most dangerous accident locations.

When comparing these 839 locations with the 800 accident locations that are currently considered as dangerous, results show that only 503 of the 800 current dangerous accident locations have a minimum estimated probability that is larger than 0.15. This indicates that 297 accident locations are currently considered as belonging to the 800 most dangerous locations (current rank between 66 and 800) while according to the ranking criterion proposed in this case study, the minimum probability for these locations is lower than expected under random conditions.

Furthermore, results of figure (3.10) and figure (3.11) show that using the lower limit of the 'confidence intervals' to select the accident locations with an estimated probability of belonging to the 800 most dangerous accident locations narrows down the number of sites from 1,431 to 839. Consequently, the use of 'confidence intervals' results in a more rigorous estimate of the most dangerous accident locations.

3.7.3 Conclusions

In this research, we used a combination of 3 different criteria, that each were studied in earlier sections, to identify and rank the accident locations. First, we only took into account the most serious injury per accident and used a valuation of casualties based on direct costs, indirect costs and validation for human suffering to give weight to the accidents. This valuation resulted in the weighing values 1_ 7_ 33 when the most severe injury respectively concerns a light, serious or deadly injury. Next, we generated probability plots, based on estimates from a hierarchical Bayes model, in order to visualize the estimated probability that a location will be ranked as dangerous.

Results showed that combining these ranking criteria will have a big impact on the selection and ranking of dangerous accident locations. In particular, when selecting the 800 most dangerous accident sites of all accident locations, 40.6% of these locations will differ from the current selection. Considering this impact quantity, we want to sensitize government to carefully choose the criteria for ranking and selecting accident locations without stating that the criterion used in this case study should be preferred to the currently used ranking method. It is up to the government to carefully decide which priorities should be stressed in the traffic safety policy. Then, the according weighing value combination can be chosen to rank and select the most dangerous accident locations.

3.8 An Optimization Framework for Injury Weights

3.8.1 Optimizing the Injury Weighing Values

In section 3.3 of this dissertation, we showed that changing the 1_ 3_ 5 weighing values that are currently used in the priority value formula for respectively a lightly injured, seriously injured and deadly injured person will have an important impact on the selection of the most dangerous accident locations. Furthermore, we explained that the choice for the values of these weight parameters are mainly a policy decision depending on the government's priorities in the traffic safety policy. Indeed, the choice between avoiding as much victims as possible or prioritizing more serious injury accidents represents different attitudes from the government towards the traffic safety problem and accordingly will have an impact on the resulting future traffic safety decisions. Government should therefore carefully decide which priorities should be stressed in the traffic safety policy. Then, the according weighing value combination can be chosen to rank and select the most dangerous accident locations.

In this section, we develop a constrained optimization model in order to automatically generate the 'optimal' weighing values for the ranking and selection problem. In general, the objective of constrained optimization models is to use mathematical or operations research procedures to find unknown variables that maximize a particular objective subject to a number of constraints using quantitative data. Accordingly, assuming that the government's objective is to select and tackle those accident locations that are primarily responsible for the human suffering on the Belgian roads, we can develop an optimization model to find the weighing values that best fit this objective.

3.8.2 Model Specifications

3.8.2.1 The optimization criterion

As explained in the introduction of this section, a first decision that needs to be taken in an optimization framework is with respect to the choice of the optimization criterion. Here, the objective is to select those accident sites that are primarily responsible for the human suffering on the Belgian roads. Analogously with the current priority value formula of the Belgian government, we take into account the number of light injuries, serious injuries and deadly injuries per accident site in order to measure the human suffering on these locations. However, now the different weighing values for each injury type are considered to be unknown variables that need to be optimized by the model to such an extent that, given a number of constraints, the sum of the human suffering on the accident sites that are selected for treatment is maximal.

This optimization problem leads to the following model specification:

Obj

$$\max Z = \sum_i (l_i L_i + s_i S_i + d_i D_i) \tag{3.9}$$

s.t.

$$l_i \leq L_{max} \tag{3.10}$$

$$s_i \leq S_{max} \tag{3.11}$$

$$d_i \leq D_{max} \tag{3.12}$$

$$L_{max} = 1 \tag{3.13}$$

$$S_{max} \leq C_1 \tag{3.14}$$

$$D_{max} \leq C_2 \tag{3.15}$$

$$L_{max} + S_{max} + D_{max} = C_3 \tag{3.16}$$

$$L_{max} \leq S_{max} \tag{3.17}$$

$$S_{max} \leq D_{max} \tag{3.18}$$

$$l_i \leq s_i \tag{3.19}$$

$$s_i \leq d_i \tag{3.20}$$

$$\sum X_i = n \tag{3.21}$$

$$l_i + s_i + d_i \leq n X_i \tag{3.22}$$

This model will be explained into more detail in the following paragraphs.

3.8.2.2 The Objective Function

On the one hand, the objective function (3.9) contains the number of lightly injured persons L_i, the number of seriously injured persons S_i and the number of deadly injured persons D_i at a location i (i = 1,..., n) that are given from the data. On the other hand, this function contains the unknown variables l_i, s_i and d_i which represent the weighing values for respectively a lightly injured, seriously injured and deadly injured person. For each location i these variables will be optimized by the model to maximize the total amount of suffering Z on the number of locations n that will be selected for treatment.

3.8.2.3 The constraints

As explained in the previous paragraph, the optimization model will optimize the unknown weighing values in order to maximize the total amount of human suffering on the selected locations. If no constraints were added to this objective function, the model would simply set these weights to an infinite high value in order to achieve the highest possible value for the objective function. However, this not the case. More specifically, the constraints (3.10), (3.11) and (3.12) state that the different weighing values per location should be less or equal than the maximum weighing value for these injury types. These maximum weighing values are also optimized by the model but are given an upper bound in the constraints (3.13), (3.14) and (3.15) by means of the user defined parameter values C_1 and C_2. By setting the parameter value C_3 in equation (3.16) smaller than the sum of C_1 and C_2, we prevent the model from setting the maximum weighing values equal to the upper bound values C_1 and C_2 and force the model to optimize the maximum weighing values depending on the data. Note that the maximum weighing value for the lightly injured persons is set at '1'. The upper bound values C_1 and C_2 can then be chosen in proportion with this value for the light injuries in function of ethical or economic viewpoints from the government (see section 3.3). In this context, equations (3.17), (3.18) and (3.19), (3.20) force the optimization model to respect the fact that most traffic safety policies will valuate a deadly injury higher as a serious injury, and, in turn, a serious injury higher as a lightly injury. Finally, in equations (3.21) and (3.22) a boolean variable per location X_i is used to, analogously with the current selection of dangerous accident sites, select from the 1,014 accident sites a subset of 800 (n = 800) accident locations (for which X_i=1) that result in the maximum human suffering on the Belgian roads using the optimized weighing values.

3.8.2.4 Mixed Integer Programming

In this research, we use a General Algebraic Modeling System (GAMS) to optimize the weighing values. Since various types of models can be solved with GAMS, the type of model must be declared before it is solved. Here, we deal with a 'Mixed Integer Programming Model'. This model can contain both discrete (binary or integer) variables and non discrete variables. Additionally, the discrete variables must assume integer variables between their bounds. Indeed, we specified a discrete binary variable X_i for each accident location i, which can only assume the values '1' (if the accident site is selected) or '0' (if the accident site is not selected). Furthermore, for the different weighing values we specified the positive variables l_i, s_i, d_i, L_{max}, S_{max} and D_{max}, which are non discrete variables and can assume any value between 0 and $+\infty$.

In this research, we will focus on the 1,014 accident locations that are currently considered as dangerous in Flanders. More specifically, for these accident sites we will use the mixed integer programming model in order to select the 800 accident locations and the corresponding weights that maximize the human suffering on the Belgian roads.

3.8.3 Empirical Results

As explained in the introduction of this chapter, the currently used weighing values by the Flemish government to select and treat the most dangerous accident sites are 1_ 3_ 5 for respectively a light, serious and fatal injury. Assuming these values reflect the government's attitude and priorities in the traffic safety policy, we use the optimization model to check whether these weighing values indeed select the accident sites that are primarily responsible for the human suffering on the Belgian roads.

Therefore, respecting the magnitude of these different weighing values and the relation between them, we set the upper bound values C_1 and C_2 respectively at 5 and 10. Next, we set the value for C_3 at 9, which is exactly the sum of the current weighing values (1+ 3+ 5). Given these maximum weighing values of 1_ 5_ 10 and the constraint that the sum of the weighing values should equal the sum of the currently used weighing values, we calculate the optimal weighing values that maximize the total amount of human suffering for the selected sites:

$$L_{max} = 1, S_{max} \leq 5, D_{max} \leq 10$$

$$L_{max} + S_{max} + D_{max} = 9$$

These model specifications give the following optimal weighing values for selecting the most dangerous sites:

$$S_{max} = 4,\ D_{max} = 4$$
$$l_i = 1,\ s_i = 4 \text{ or } 1,\ d_i = 4$$

These results show that, in order to maximize the human suffering on the Belgian roads, instead of choosing the 1_ 3_ 5 weighing value combination or maximizing the weight for a deadly injury, the model chooses to maximize the weight for a seriously injured person. Indeed, given the constraints $s_i \leq d_i$ and $L_{max} + S_{max} + D_{max} = 9$, the maximum value the model can assign to S_{max} is 4. Correspondingly, the value for D_{max} then equals 4. For the 800 locations that are selected the values for l_i, s_i and d_i then respectively equal the values 1_ 4_ 4. Note however, when the number of seriously injured persons at the location is 0, the model assigns a value of 1 to s_i, but this has no impact on the value of the objective function, nor on the optimal weighing value for d_i.

In order to validate these results and test the impact of the constant value C_3 on the optimal solution of the model, we solved two similar models using different values for C_3. More specifically, we first solved the model using the constraint $L_{max} + S_{max} + D_{max} = 10$. This resulted in the following maximum and optimal weighing values:

$$S_{max} = 4.5,\ D_{max} = 4.5$$
$$l_i = 1,\ s_i = 4.5 \text{ or } 1,\ d_i = 4.5$$

For the second alternative model we used the constraint $L_{max} + S_{max} + D_{max} = 8$. This resulted in the following weighing values:

$$S_{max} = 3.5,\ D_{max} = 3.5$$
$$l_i = 1,\ s_i = 3.5 \text{ or } 1,\ d_i = 3.5$$

These figures show that obviously the value for C_3 will influence the values for the maximum weighing values S_{max} and D_{max} and accordingly the optimal values for s_i and d_i. However, regardless of the value of C_3 and regardless of the higher value for C_2, the model will maximize the weight for a serious injury.

Note that, if we omit from the model the constraints that $S_{max} \leq D_{max}$ and $s_i \leq d_i$, for the first alternative model, the value for S_{max} will increase to 5 while the value for D_{max} will decrease to 4. Similarly, for the second model, the value for S_{max} will increase to 4 while the value for D_{max} will decrease to 3. These figures confirm the result that the model tends to maximize the values for S_{max} and S_i. However, these results do not stroke with government's attitude to higher value a deadly injury than an serious injury.

These results can be explained by the nature of the accident data. More specifically, for the 1,014 accident locations, on average per location 0.4 deadly injured casualties, 2.9 seriously injured casualties and 14.2 lightly injured casualties occurred. This means that the number of seriously injured persons is on average almost 8 times higher than the number of deadly injured persons per location. Since in the objective function these numbers L_i, S_i and D_i will be multiplied by the optimal weights l_i, s_i, d_i, it can be easily understood why the optimization model will maximize the value for s_i instead of d_i.

3.8.4 Conclusions

In this research, we developed a constrained optimization model in order to automatically generate the 'optimal' weighing values for the ranking and selection problem. In particular, assuming that the government's objective is to select and tackle those accident locations that are primarily responsible for the human suffering on the Belgian roads, we used a mixed integer programming model to find the weighing values that best fit this objective.

Results showed that in order to select the accident sites that maximize the human suffering on the Belgian roads, instead of choosing the 1_ 3_ 5 weighing value combination or maximizing the weight for a deadly injury, the model chooses to maximize the weight for a seriously injured person. This can be explained by the relatively high number of serious injuries per accident locations.

Accordingly, by choosing the weighing value combination 1_ 3_ 5, the Flemish government does express its preference to higher value the human suffering of a deadly injury than a serious or light injury. However, this weighing value combination does not select the accident locations that are mainly responsible for the total amount of human suffering on the roads. Therefore, in order to select the accident locations that maximize the human suffering on the roads, one should not only take into account the valuation of injury types from an economic or ethical point of view (represented by

C_1 and C_2) but also the relation between the number of injury types at each location.

Note that, in this research, when optimizing for human suffering to maximize the benefit of reducing that suffering, we are really optimizing 'random' suffering and 'systematic' suffering, the former deriving from the random fluctuation of crashes observed at sites, the latter coming from true deficiencies at sites, both of which lead to crashes and suffering. Accordingly, in order to optimize for true' risk and not merely observed risk' we should know the underlying true safety of the locations. However, as explained in section 3.2.3, this is not the case. One other possible solution to account for this random suffering could be to take into account the estimated instead of the observed number of injuries to measure the human suffering on the locations. This should be explored in further research.

3.9 Discussion and limitations

In this chapter, a sensitivity analysis is performed on the currently used method to rank and select dangerous accident locations in Flanders. This study is based on the same data used to select and rank the 1,014 accident sites that are currently considered as dangerous by the Flemish government. However, since in this research, no simulated accident data were used, we can only estimate the effects on the current ranking of accident locations and the locations that are currently ranked as most dangerous, without claiming that these sites are indeed the truly most dangerous sites. Furthermore, in order to quantify the effects of changing the ranking and selection criteria of dangerous accident locations, we use the percentage deviation value. This measure allows comparing the elements of two data sets of equal size containing different locations. However, this measure only gives information about the number of locations that do not appear in two ranked data sets and does not take into account internal shifts in the ranking position of the common accident locations. Accordingly, it is possible for two ranking methods to yield the same value for the percentage deviation value but be very different in terms of their ranking. Two methods might identify the same top 10 most dangerous sites, but rank them very differently and also rank the remaining sites very differently. The degree of 'how far off' sites are ranked compared to their true deserved ranking is not considered. As explained in the introduction of this dissertation, this acquires the knowledge of the true underlying safety of each location in order to know the true rank to compare with the assigned rank via the alternative ranking methods.

First of all, this research showed that changing the 1_ 3_ 5 weighing values that are currently used for respectively a lightly injured, seriously injured and deadly injured person will have an important impact on the selection of the most dangerous accident locations. However, the choice for the values of these weight parameters are mainly a policy decision depending on the government's priorities in the traffic safety policy. Therefore, no conclusions were made concerning which weighing criterion should be preferred. This was not the objective of this research and will require additional data and in depth analysis of the accident locations. In contrast, with this research we want to make the government aware that choosing different weighing values will greatly influence the selection of hazardous accident locations. This might seem straightforward, however, in Flanders no such research was yet conducted. Accordingly, the weighing values used in this sensitivity analysis are just examples and therefore no statement is made that one of these four alternatives is the optimal weighing scheme.

Secondly, analysis showed that giving weight to the severity of the accident instead of to all the injured occupants of the vehicle will also have a great impact on the selection and ranking of dangerous accident locations. By giving weight to the severity of the accident we can correct for the bias in the priority score that occurs when the number of occupants of the vehicles are subject to coincidence. However, in some cases (e.g. discotheques, entertainment centers), it can be reasoned that the number of occupants, and accordingly the number of injured persons, is not a coincidence but more likely a trend. For these locations, correcting for the number of passengers would not be advisable since the number of injuries that appear at these locations are inherent to the locations characteristics.

Next, we used a multivariate hierarchical Bayes approach for ranking the accident sites, taking into account the number of accidents, the number of fatalities and the number of light and severely injured casualties. This research showed that the use of Bayesian estimation values instead of historic count data to rank the accident locations will take into account the problem of random variation in accident counts and will also have an important effect on the selection of the most dangerous accident locations. Note that in this research, in order to become an estimate for the long term accident mean per location, the Hierarchical Bayes model was fitted by applying it to the empirical accident frequency, available from the accident data. Another interesting approach for estimating the number of accidents could be to estimate the number of accidents as a function of various explanatory variables by means of a multivariate model (see also section 3.5.2). Then Empirical Bayes could be applied to predict the expected number of accidents for each locations. However, in this research, no such accident prediction model was developed. The reason for this is that, in this dissertation, we want to investigate the strengths and weaknesses of the currently used method to identify and rank dangerous accident locations in Flanders. As explained in section 2.3, the Flemish government first draws up a list of dangerous accident sites, based on the number of accidents and injuries at each location. Information on the explanatory variables for this number of accidents, such as traffic volume, specific variables describing road design, traffic control and the surroundings of the road are not available when drawing up this list of black spots, and are only collected when developing a conceptual solution for enhancing the level of traffic safety as these locations. Accordingly, a Hierarchical Bayes model was developed based on only the empirical accident frequency to estimate the expected number of accidents at each location. This information is taken into account by the model in the form of the likelihood. Furthermore, prior knowledge is included to estimate the expected number of accidents for each location. This method can handle the uncertainty and great variability of accident data and produce a probabilistic ranking of the accident locations.

Additionally, Bayesian ranking plots can be used to visualize the estimated probability that a location will be ranked as belonging to the most dangerous locations. These probability plots can provide policy makers with a scientific instrument with intuitive appeal to select dangerous road locations on a statistically sound basis. In this research, this implied that we used this technique to estimate the probability that a location will be selected as one of the 800 most dangerous sites, since for budgetary reasons this is the number of sites that the Flemish government will redesign to enhance traffic safety. Results showed, however, that more than 800 sites qualify to be considered as one of the 800 most dangerous locations. Therefore, we want to sensitize government that that the choice for tackling only 800 accident locations seems somewhat arbitrary. However, no statement was made in this research on how many sites should be tackled by the government since this, of course, will depend on the budget that is available to improve traffic safety.

Furthermore, we used a combination of the three different criteria that were discussed above, to identify and rank the accident locations. First, we only took into account the most serious injury per accident. Secondly, we used a valuation of casualties based on direct costs, indirect costs and validation for human suffering to give weight to the accidents. This valuation resulted in the weighing values 1_ 7_ 33 when the most severe injury respectively concerns a light, serious or deadly injury. Based on estimates from a hierarchical Bayes model, we generated probability plots, in order to visualize the estimated probability that a location will be ranked as dangerous. Results showed that combining these ranking criteria will have a big impact on the selection and ranking of dangerous accident locations. Considering this impact quantity, we want to sensitize government to carefully choose the criteria for ranking and selecting accident locations without stating that the criterion used in this paper should be preferred to the currently used ranking method. Indeed, the 1_ 7_ 33 weighing values seem a reasonable and scientifically sound alternative for the currently used 1_ 3_ 5 weighing combination, but as explained before, it is up to the government to decide which priorities should be stressed in their traffic safety policy. Then, the according weighing value combination can be chosen to rank and select the most dangerous accident locations.

Finally, we used mixed integer programming in order to automatically generate the ‘optimal’ weighing values for the 800 accident locations that are primarily responsible for the human suffering on the Belgian roads. Results showed that, given the optimization framework, instead of choosing the 1_ 3_ 5 weighing value combination, the model chooses to maximize the weight for a seriously injured person (1_ 4_ 4). This can be explained by the relatively high number of serious injuries per accident locations. Therefore, in order to select the accident locations that maximize the human suffering on the roads, the government should not only take into account the valuation of injury types from an economic or ethical point of view (represented by C_1 and C_2) but also the relation between the number of injury types at each location.

Chapter 4

Profiling Hazardous Accident Locations

In this chapter, dangerous accident locations are profiled in terms of accident related data and location characteristics using different data mining and statistical techniques. Furthermore, these techniques will be used to develop models, which will provide new insights into the criteria that aim to explain the occurrence of a traffic accident[1].

[1] Parts of this chapter have been published as follows:

- Geurts, K., Thomas I. and G. Wets (2005). Understanding accidents in black zones using frequent item sets. Accident analysis and prevention 37 (4). pp.787-799
- Geurts K., Wets G., Brijs T., and Vanhoof K. (2004), Profiling of high frequency accident locations by use of association rules. In Journal of Transportation Research Board, Vol. 1840, pp. 123-130. ISBN: 0309085810.
- Geurts K., Wets G., Brijs T. and Vanhoof K. (2003), Clustering and profiling traffic roads by means of accident data. Electronic proceedings of the European Transport Conference, Strasbourg, France,October 8-10, 16 pp. ISBN 0-86050-342-9
- Geurts, K. Wets, G. and Brijs, T. (2003). Profiling high frequency accident locations using association rules. Electronic Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington, January 12-16, USA, 18 pp.
- Geurts K., Wets G., Brijs T. and Vanhoof K. (2002), The Use of Rule Based Knowledge Discovery Techniques to Profile Black Spots, the 6th Design and Decision Support Systems in Architecture and Urban Planning Conference, Ellecom, The Netherlands, July 7-10, pp 15.

4.1 Data Mining vs. Statistical Models

As explained in figure 1.2, after the hazardous sites are identified, it is necessary to carefully examine the nature of the safety problem at the sites with a view to identifying whether and how those problems can be dealt with through road or traffic remedial measures. In this context, Kononov and Janson (2002) explain that without being able to properly and systematically relate accident frequency and severity to roadway geometrics, traffic control devices, roadside features, roadway condition, driver behavior or vehicle type it is not possible to develop effective countermeasures. Therefore, in the past, statistical models have been widely used on accident data to analyze road crashes in order to explain the relationship between crash involvement and traffic on the one hand and geometric and environmental factors on the other hand (Lee et al. (2002)). For example, Kononov and Janson (2002) have developed a methodology to detect accident patterns at an intersection, which suggests a presence of an element or elements in the roadway environment which triggered a deviation from a random statistical process in the direction of reduced safety. However, Chen and Jovanis (2002) indicate that not only the main effects of driver, vehicle, roadway and environmental factors should be analyzed, interactions between factors are also very likely to be significant. The authors demonstrate that the large number of potentially important factors, combined with the complex nature of crash etiology and injury outcome present certain challenges when using classic statistical analysis on data sets with large dimensions such as an exponential increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables. Furthermore, a large number of factors need to be selected and a comprehensive but feasible set of main factors and interactions need to be specified for testing in statistical models.

This is where data mining comes into play. Data mining can be defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996). From a statistical perspective it can be viewed as a computer automated exploratory data analysis of (usually) large complex data sets (Friedman, 1997). More specifically, data mining attacks such problems as obtaining efficient summaries of large amounts of data, identifying interesting structures and relationships within a data set, and using a set of previously observed data to construct predictors of future observations. Statisticians have well established techniques for attacking all of these problems as well. Indeed, a statistician might argue that data mining is not much more than the scaling up of conventional statistical methods to massive data sets, in effect a large scale 'data engineering' effort (Smyth, 2001). However, the problems and methods of data mining have some distinct features of their own. Not only can data sets be much larger than in statistics and are data analyses on a correspondingly larger scale, there are also differences of emphasis in the approach to modeling: compared with statistics, data mining pays less attention to the large-scale asymptotic properties of its inferences and more to the general philosophy of 'learning', including consideration of the complexity of models and the computations they require (Hosking et al., 1997). Furthermore, data mining has tackled problems such as what to do in situations where the number of variables is so large that looking at all pairs of variables is computationally infeasible (Mannilla, 2000). Additionally, opposite to statistics, data mining is typically secondary data analysis: the data has been collected for some other purpose than for answering a specific data analytical question.

In literature, some examples of the use of data mining in road accidents analysis can be found. For example, clustering techniques are used to discover frequent patterns in accident data (see e.g. Ljubic et al., 2002). Additionally the data mining technique of rule induction can be used to identify rules sets representing interesting subgroups in accident data (see e.g. Kavsek et al., 2002). Furthermore, decision trees (see e.g. Strnad et al., 1998) and neural networks (see e.g. Mussone et al., 1999) are used to model and analyze road accidents. Finally, spatial data mining (see e.g. Chelghoum and Zeitouni, 2004) can be applied.

4.2 The KDD Process

As explained in the introduction, data mining is used to discover patterns and relationships in data, with an emphasis on large, observational data bases (Friedman (1997)). According to Fayyad et al. (1996) it can be considered as a separate step of the 'knowledge discovery in databases' (KDD) process (see figure 4.1).

This KDD process refers to the overall process of discovering useful knowledge from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge and proper interpretation of the results of mining are essential to ensure that useful knowledge is derived from the data. Blind application of data mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

The first step of the KDD process involves the collection of the data. In this

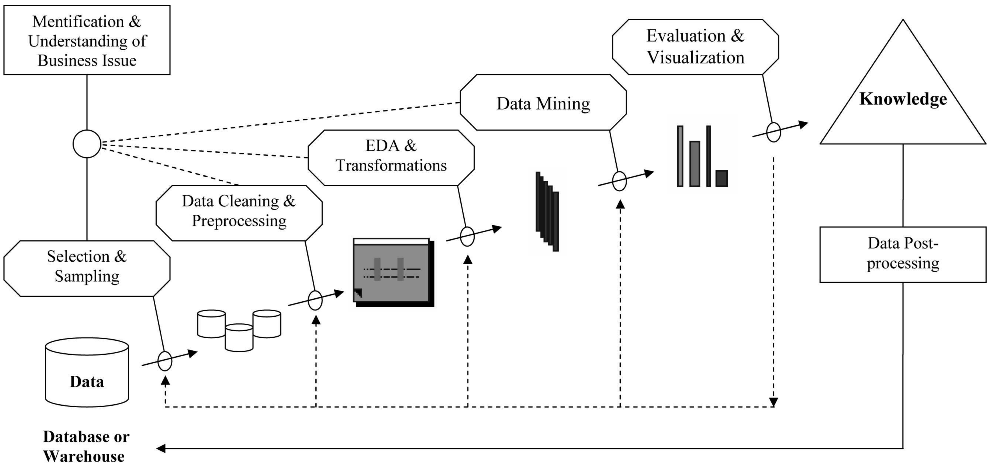

Figure 4.1: The KDD process (Source: Fayyad et al., 1996)

step, the data are selected or segmented according to some criteria. Next, in the preprocessing step of the KDD process, the data that are collected in the previous data mining step are converted into usable data in order to perform pattern discovery techniques. Irregularities such as data inconsistencies, missing values, redundant variables and double counts will be tracked, listed and finally removed from the data. Additionally, some new items will be created based on existing information of the data set. Also the data is reconfigured to ensure a consistent format, as there is a possibility of inconsistent formats because the data are often drawn from different sources. In the transformation step, the data are reduced and transformed so that they can be interpreted and analyzed by the data mining algorithm. Examples of such adjustments are changing continuous attribute values into discrete attribute values and changing items with nominal attribute values into items with discrete attribute values. A more detailed description of these different data preparation steps for data mining can be found in Pyle (1999).

In the data mining step, the actual analysis of the data takes place. By means of different software tools the data-mining algorithm will look for patterns in the data. Depending on the objectives of the research, different data mining techniques can be implemented. In general, the four following data mining tasks can be discerned (Berry and Linoff, 1997): classification, estimation, affinity grouping and clustering. These data mining tasks can be classified into two major categories, i.e. prediction and description. Indeed, classification and estimation tasks can be seen as prediction tasks:

the user wants to predict the (discrete or continuous) value of an unknown attribute (e.g. using decision trees). On the other hand, affinity grouping and clustering can be seen as description tasks: it is the objective of the analyst to gain insight into the underlying relationships that exist between attributes or instances in the database. However, these two categories are not mutually exclusive in the sense that some techniques can be used for both purposes (Fayyad et al., 1996).

In this research, we will use some data mining techniques of affinity grouping (by means of the association algorithm) and clustering to describe the accident data. This will be explained in the following section. However, the choice for one data mining technique or another is not always obvious. In fact, this choice is largely dependent on the situation. Each technique has its strengths and weaknesses in terms of representation language, classification power, descriptive abilities and expert knowledge required. Therefore, the analyst has to evaluate what kind of problem he is faced with to choose for the appropriate technique. For example, the association algorithm is able to identify all the accident circumstances that frequently occur together. However, no explanation about the causality of these accident patterns is given. Furthermore, to evaluate and interpret the interestingness of the results domain knowledge or the use of additional statistical techniques are essential.

This occurs in the last step of the KDD process: the patterns that are identified by the data mining algorithm are evaluated and interpreted into knowledge which can than be used to support human decision making.

4.3 Methods and Techniques

In this section, the two techniques that are used in this research to profile hazardous crash locations are discussed: the association algorithm and model based clustering.

4.3.1 Association Algorithm

4.3.1.1 Frequent Item Sets

In this study, an association algorithm is used to obtain a descriptive analysis of the accident locations. This data mining technique fulfils the task of affinity grouping and was first introduced by Agrawal et al. (1993). It can be used to efficiently search for interesting information in large amounts of data. Since its introduction, the task of association mining has received a great deal of attention in the data mining community. Its direct applicability to business problems together with their inherent understandability, even for non-data mining experts, made the association algorithm a popular mining method. Today mining associations is still one of the most popular pattern discovery methods in KDD (Hipp et al., 2000). For example, this technique has been adopted in different contexts, such as retailing (Brijs, 2002), cross-selling (Anand et al., 1997), finding co-occurring medical tests from a health insurance information system (Viveros et al., 1996), reducing fall-out in telecommunications systems (Ali et al., 1997) and identifying latently dissatisfied customers (Bloemer et al., 2002).

In particular, the association algorithm produces frequent item sets describing underlying patterns in data (Agrawal et al., 1993):

- Let $I = i_1, i_2, ..., i_k$ be a set of literals, called items.

  Example: The accident characteristics in an accident database.

- Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$.

  Example: A data set D with 4 accidents, where each accident T contains, a set of items X, representing different accident characteristics I.

  1. Rain, intersection, traffic lights
  2. Rain, intersection, traffic signs
  3. Normal weather, zebra crossing, pedestrian
  4. Rain, bicycle, cycle track.

In contrast to predictive accident models, the strength of the association algorithm lies within the identification of item sets that frequently occur together. Moreover, the association algorithm is able to generate all accident patterns taking into consideration the minimum support value. The support of an item set indicates how frequent that combination of items or accident characteristics occurs in the data. The higher the support of the item set, the more prevalent the item set is. It is obvious that we are especially interested in item sets that have a support greater than the user-specified minimum support (minsup). These items are considered to be ‘frequent’ item sets.

- The item set X has support s in D, if s% of the accidents in D contains X.

  Example: The support expresses the fraction of accidents in D that contain all the items in the item set. Thus a support of 20% means that the accident characteristics in the item set occur in 20% of all accidents in the data set.

A typical approach (Agrawal et al., 1996) to discover all frequent item sets is to use the insight that all subsets of a frequent set must also be frequent (also known as the 'downward closure' principle). This insight simplifies the discovery of all frequent sets considerably since not all item combinations need to be tested but only those for which every subset was previously found to be frequent. This dramatically improves the efficiency of the algorithm, i.e. first find all frequent sets of size 1 by reading the data once and recording the number of times each item A occurs. Then, form candidate sets of size 2 by taking all pairs {B, C} of items such that {B} and {C} both are frequent. The frequency of the candidate sets is again evaluated against the database. Once frequent sets of size 2 are known, candidate sets of size 3 can be formed; these are sets {B, C, D} such that {B, C}, {B, D} and {C, D} are all frequent. In other words, to determine the frequent item sets of size n, form candidate item sets of size n using only frequent item sets of size n-1. Finally, evaluate the frequency of these candidate items sets of size n against the database. This process is continued until no more candidate sets can be formed.

Suppose we set the minimum support value (minsup) = 50%. This means that the accident characteristics should occur in at least 50% of all the accidents in order to be considered as frequent. For the above example, this would lead to the following results:

- Example:

  1. Frequent item sets of size 1:
     s(rain) = 3/4 (75%), s(intersection) = 2/4 (50%)
  2. Frequent item sets of size 2:
     s(rain, intersection) = 2/4 (50%)

4.3.1.2 Association Rules

Based on the frequent sets, the association algorithm can produce a set of rules describing underlying patterns in the data by means of the support parameter and the confidence parameter. Agrawal et al. (1993) provided the following formal description of this technique: An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$.

- The rule $X \Rightarrow Y$ holds in the data set D with confidence c, if $c\%$ of the accidents in D that contains X also contains Y.

  In statistical terms, the confidence is an estimator for the conditional probability of Y given X, i.e. $\mathrm{P}(Y|X)$ and it can be calculated as $\mathrm{s}(X \cup Y)/\mathrm{s}(X)$.

Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (minsup) and minimum confidence (minconf). Accordingly, generating association rules involves looking for frequent item sets in the data. Indeed, the support of the rule $X \Rightarrow Y$ equals the frequency of the item set $\{X, Y\}$. Thus by looking for frequent item sets, we can determine the support of each rule (Manilla, 1997). The problem of discovering association rules can therefore be decomposed into two sub-problems:

1. Generating all item sets that have a support higher than the user-defined minsup. These item sets are called frequent item sets.

2. Use this collection of frequent sets to generate the rules that have confidence higher than the user-defined minimum confidence.

Once all frequent sets are known, finding association rules is easy. Namely, for each frequent set X and each $Y \in X$, verify whether the rule $X \backslash \{Y\} \Rightarrow Y$ has sufficiently high confidence. The given algorithm has to read the database at most K+1 times, where K is the size of the largest frequent item set.

Suppose we are interested in the association rules that are frequent with minimum confidence value = 70%. For the above example, this would lead to the following results:

- Example:
  Candidate Rules from frequent set(rain, intersection):
  c(rain $\Rightarrow$ intersection) = 2/3 (66.6%)$<$ 70%

c(intersection ⇒ rain) = 2/2 (100%)> 70%
→ This is the only reliable association rule for this data set.

4.3.1.3 Interesting Patterns

The association algorithm generates all item sets that have support higher than the minimum support value. However, a large number of the generated item sets and association rules will be trivial and give no additional information towards the traffic accident analysis. An example of such a trivial rule is:

- Wet road surface ⇒ Rain (sup= 19.91%, conf= 68.21%)

The confidence value shows that the probability of observing this kind of weather is 68.21% if the antecedent is true (wet), meaning that 68.21% of the days that an accident happens on a road with a wet surface it will be raining. The support of the rule indicates that 19,91% of all the accidents that occurred, happened on a wet road surface while it was raining. However, since it is obvious that even when no accidents happen the probability for rain will be higher with a wet road surface, this rule does not give an additional value towards a better understanding of the circumstances in which these accidents have happened. Therefore, this rule will be removed from the association rules set.

Therefore, a filter is needed to post-process the discovered item sets to retain the interesting ones. In literature several filtering procedures for association patterns exist (see e.g. Tan et al., 2002). In this research, we will use the following filter methods.

Lift
A first, more formal method (Brin et al., 1997) to distinguish trivial from non trivial rules is assessing the dependence between the items in the item set by means of the lift value (L):

$$L = \frac{s(A,B)}{s(A) * s(B)} \tag{4.1}$$

The nominator in equation (4.1), $s(A,B)$, measures the observed frequency of the co-occurrence of the items A and B. The denominator $s(A)*s(B)$ measures the expected frequency of the co-occurrence of the two items under the assumption of conditional independence. The more this ratio differs from 1, the stronger the dependence. Table (4.1) illustrates the three possible outcomes for the lift value and their associated interpretation for the dependence between the items.

Table 4.1: Interpretation of Lift

Outcome	**Interpretation**
$+\propto > L > 1$	Positive interdependence effects between X and Y
$L = 1$	Conditional independence effects between X and Y
$0 < L < 1$	Negative interdependence effects between X and Y

For the item set (rain, intersection), the lift value is calculated as follows:

- Example:
  Lift (rain, crossroad) = $\frac{s(rain,crossroad)}{s(rain)*s(crossroad)} = \frac{(2/4)}{(3/4)*(2/4)} = 4/3\ (>1)$
  → small positive interdependence between rain and crossroad.

Interest

Besides ranking the item sets on their lift value we can use a second measure, i.e. the interest measure, to limit the accident patterns to only the discriminating or useful ones (Anand et al. (1997)):

$$\text{Int} = \frac{s_1 - s_2}{max\{s_1, s_2\}} \tag{4.2}$$

This interest measure is based on the deviation in support values of the frequent item sets discovered for two different groups (for example, two groups of accident locations). The nominator $s_1 - s_2$ measures the difference in support for the accident characteristics group 1 (s_1) and group 2 (s_2). The expression max is called the normalizing factor as it normalizes the interestingness measure onto the scale [-1, 1].

Suppose the item set (rain, crossroad) is also frequent in a second data set with: s_2(rain, crossroad) = 3/4 (75%). The interest value can than be calculated as follows:

- Example:
  Interest = $\frac{s_1(rain,crossroad)-s_2(rain,crossroad)}{\max(s_1,s_2)} = \frac{(2/4)-(3/4)}{(3/4)}$ = -1/3 (<1)

  → this value is close to '0' indicating that this item set is not very discriminating between the two data sets.

Statistical Rule Significance

In the case of association rules, besides the lift and interest measure, we can use a third measure to discern trivial from non trivial rules: the statistical rule significance

(T). More specifically, the statistical significance of a rule is the validity of a rule, based on the influence of statistical dependency between the rule body (antecedent) and the rule head (consequent). T is determined using the Chi-square test for statistical independence (Brin et al., 1998), which is a widely used test for identifying propositional rule independence (see equation 4.3). For example Liu et al. (1999) used a chi-square test to decide whether the antecedent of a rule is independent from its consequent or not.

The following example illustrates the calculation of the χ^2 value. Suppose the following situation. Among 5,000 road accidents:

- 3,000 occurred on an intersection
- 3,750 occurred in the rain
- 2,000 occurred on an intersection in the rain

Table 4.2 shows the contingency table that can be drived from this example.

The χ^2 test for independence is calculated as follows, with O_{XY} the observed frequency and E_{XY} the expected frequency (by multiplying the row and column total divided by the grand total):

$$\chi^2 = \sum_{X,Y} \frac{(O_{XY} - E_{XY})^2}{E_{XY}} \tag{4.3}$$

If the χ^2 value < critical χ^2 value, there is statistical independency between the rule body and the rule head and the statistical significance (T) will be neutral. If the χ^2 value > critical χ^2 value, there is statistical dependency between the rule body and the rule head. Depending on the relationship between the observed (O_{XY}) and the expected (E_{XY}) frequencies, the algorithm determines whether the statistical significance (T) is negative or positive. For this example χ^2 will be calculated as

Table 4.2: Contingency Table

	Rain	**No rain**	**Total**
Intersection	2,000	1,000	3,000
No intersection	1,750	150	2,000
Total	3,750	1,150	5,000

Table 4.3: Interpretation of Statistical Rule Significance

Outcome	**Interpretation**
T < 0	- Item Set X has a negative influence on the occurrence of item set Y - Given item set X, item set Y occurs less frequently than expected - Lift between 0 and 1 - Valid rule (T = -)
T is neutral	- Lift = 1: Item Sets X and Y are statistically independent → Rule gives no extra information - Lift $\neq$ 1: Rule has failed the χ^2 test → Rule is not valid
T > 0	Item Set X has a positive influence on the occurrence of item set Y - Given item set X, item set Y occurs more frequently than expected - Lift > 1 - Valid rule (T = +)

follows:

$$\begin{aligned}
\chi^2 &= \frac{(2{,}000 - \frac{3{,}000*3{,}750}{5{,}000})^2}{\frac{3{,}000*3{,}750}{5{,}000}} + \frac{(1{,}000 - \frac{3{,}000*1{,}150}{5{,}000})^2}{\frac{3{,}000*1{,}150}{5{,}000}} \\
&\quad + \frac{(1{,}750 - \frac{2{,}000*3{,}750}{5{,}000})^2}{\frac{2{,}000*3{,}750}{5{,}000}} + \frac{(150 - \frac{2{,}000*1{,}150}{5{,}000})^2}{\frac{2{,}000*1{,}150}{5{,}000}} \\
&= 417.63 >> 3.84
\end{aligned}$$

For the *p*-value of 0.05 with one degree of freedom, the cutoff value equals 3.84. Consequently, the accident characteristics rain and intersection can be considered as highly interdependent at the 95 % confidence level.

Table (4.3) gives an overview for the possible outcomes of the statistical rule significance test (T) and indicates its relation with the lift value of the rule.

4.3.2 Model Based Clustering

Cameron (1997) indicates that clustering methods are an important tool when analyzing traffic accidents as these methods are able to identify groups of road users, vehicles and road clusters which would be suitable targets for countermeasures.

In this dissertation, we will adopt a more statistically based approach to clustering, known as latent class clustering, or also called model-based clustering or finite mixture modeling, to cluster traffic roads into distinct groups based on their similar accident frequencies. More specifically, the observed accident frequencies are assumed to originate from a mixture of density distributions for which the parameters of the distribution, the size and the number of clusters are unknown. It is the objective of model based clustering to 'unmix' the distributions and to find the optimal parameters of the distributions and the number and size of the clusters, given the underlying data (McLachlan and Peel, 2000).

It is not within the scope of this dissertation to provide an exhaustive overview of the domain of latent class cluster analysis, but rather to focus on the concepts and techniques that are most relevant for the development of the cluster model in this dissertation. This section provides an overview of the general formulation of the latent class cluster model and is drawn from Brijs (2002), who refers to some state-of-the-art books and review articles in this domain.

4.3.2.1 Modeling Accident Rates with Poisson Distribution

Since we do not know exactly what causes traffic accidents to happen, the approach is based on the idea of modeling the accident frequency as a Poisson-distributed random variable Y (see also equation (3.3)). In general, the Poisson random variable $Y_i(t)$ represents the number of occurrences of a rare event in a time interval (t) and is, therefore, well suited for modeling the number of accidents at location i (Nag et al., 2002).

Formula (4.4) shows that we are given a number of locations (i =1,..., n) on which the discrete random variable Y_i is Poisson distributed and measured over a certain period of time (t), where y_i = 0,1,2,...(i.e. accident rate) and the rate parameter $\lambda > 0$:

$$Poi(Y_i(t) = y_i \mid (\lambda t)) = \frac{e^{-(\lambda t)}\ (\lambda t)^{y_i}}{y_i!} \tag{4.4}$$

The mean and the variance of the Poisson distribution are $\mathrm{E}(Y) = \lambda t$ and $\mathrm{Var}(Y) = \lambda t$, respectively. The fact that the mean and the variance of the Poisson distribution are identical is, however, too restrictive in many applications where the variance of the data may exceed the mean (Cameron and Trivedi, 1986). This situation is called 'overdispersion' (McCullagh and Nelder, 1989) and may be due to heterogeneity in the mean event rate of the Poisson parameter λ across the sample. Solutions to the problem of overdispersion therefore involve accommodating for the heterogeneity in the model. In this research, we will adopt the finite mixture specification.

4.3.2.2 The Finite Mixture Specification

The finite mixture specification assumes that the underlying distribution of the Poisson parameter λ over the population can be approximated by a finite number of support points (Wedel et al., 1993), which in the context of this study represent different clusters or latent classes of accident locations in the data.

These support points and their respective probability masses can be estimated by a maximum likelihood approach. For instance, in the case of a two-cluster model, we assume that there are two support points. In other words, we assume there are two groups of locations:

- a group of roads of size p_1 whose latent accident parameter $\lambda = \theta_1$,
- a second group of roads of size $p_2 = (1 - p_1)$ whose average accident rate $\lambda = \theta_2$, where p_j are the mixing proportions with $0 < p_j < 1$ and

  $$\sum_{j=1}^{k} p_j = 1$$

  Note that the mixing proportion is the probability that a randomly selected observation belongs to the j-th cluster.

Consequently, the two cluster model can be formulated as follows:

$$\begin{aligned} P(Y_i(t) = y_i) &= P(Y_i(t) = y_i|\text{group1}).P(\text{group1}) \;+\; P(Y_i(t) = y_i|\text{group2}).P(\text{group2}) \\ &= \frac{e^{-(\theta_1 t)}\ (\theta_1 t)^{y_i}}{y_i!}.\ p_1 + \frac{e^{-(\theta_2 t)}\ (\theta_2 t)^{y_i}}{y_i!}.\ (1 - p_1) \end{aligned} \tag{4.5}$$

In general, the purpose of model-based clustering is to estimate the parameters $(p_1, ..., p_{k-1}, \theta_1, ..., \theta_k)$, with $k =$ the number of clusters, following the maximum likelihood (ML) estimation approach. This involves maximizing the loglikelihood. For the two-cluster model, the loglikelihood function is then defined as :

$$\text{LL}(p_1, \theta_1, \theta_2|\text{ data}) = \sum_{i=1}^{n} \ln\ (\frac{e^{-(\theta_1 t)}\ (\theta_1 t)^{y_i}}{y_i!}.\ p_1 + \frac{e^{-(\theta_2 t)}\ (\theta_2 t)^{y_i}}{y_i!}.\ (1 - p_1)) \tag{4.6}$$

In this research, we use a non-linear iterative fitting algorithm (nlp) to maximize the loglikelihood. To prevent the algorithm from finding a local but not a global optimum, we use multiple sets of starting values for the algorithm and we observe the evolution of the final likelihood for different restarts of the algorithm.

4.3.2.3 Determining the Number of Clusters

To decide on the number of components in a mixture model, we use the so-called information criteria to evaluate the quality of a cluster solution. Examples include AIC (Akaike information criterion), CAIC (Consistent Akaike information criterion) and BIC (Bayes information criterion) (Schwarz, 1978):

$$\begin{aligned} \text{AIC} &= -2\text{L}_k + 2.d_k && (4.7)\\ \text{BIC} &= -2\text{L}_k + \ln(n).d_k && (4.8)\\ \text{CAIC} &= -2\text{L}_k + [\ln(n)+1].\ d_k && (4.9) \end{aligned}$$

These are goodness of fit measures, which take into account model parsimony. The idea is that the increase of the likelihood of the mixture model (L_k) on a particular data set of size n, is penalized by the increased number of parameters (d_k) needed to produce this increase in fit. The smaller the criterion, the better the model in comparison with another.

4.4 Data

4.4.1 Road Accidents with Casualties

As explained in the previous chapter, a large data set of road accidents is available for analysis. More specifically, in Belgium, the data on road accidents with casualties of the last few decades are stored and updated in a flat file format by the National Institute of Statistics. These data are obtained from the Belgian ‘Analysis Form for Traffic Accidents’ that should be filled out by a police officer for each road accident that occurs on a public road involving casualties. In total, information on approximately 45 items is available for each accident. These data are a rich source of information on the circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries,...), traffic conditions (maximum speed, priority regulation,...), environmental conditions (weather, light conditions, time of the accident,...), road conditions (road surface, obstacles,...), human conditions (characteristics of the road user, fatigue, alcohol,...) and geographical conditions (location, rough physical characteristics,...).

Figure (4.2) shows the structure of these data, presented in an Access database. In particular, Appendix 1 explains the items that are included in the table ONGEVAL. These items measure the characteristics that are unique for each accident. Next, Appendix 2 gives information about the items in the table WEGGEBRUIKERS. This table describes the characteristics that are unique for each road user. This information can be linked to the corresponding accident by means of the unique accident ID. Furthermore, Appendix 3, Appendix 4, Appendix 5 and Appendix 6 give more detailed information about respectively the injured passengers (PASSAGIERS), pedestrians (VOETGANGERS), bikers (FIETSERS) and injured victims (SLACHTOFFERS). Again the information from these first three tables can be related to the accident and to the road user by combining the unique accident ID and the number of the road user. Since, per definition, a victim cannot be a road user, the information on the victims can only be related to the accident table by means of the unique accident ID. Then, Appendix 7 explains the items of the table LOCATIE. In this table, the items are listed that relate to the location characteristics of the accident. As explained in the previous chapter, in Belgium, the location of an accident is accurately known for the ‘numbered’ roads (i.e. highways, national and provincial roads linking towns). On these roads, every hectometer there is a stone marker. Therefore, every accident

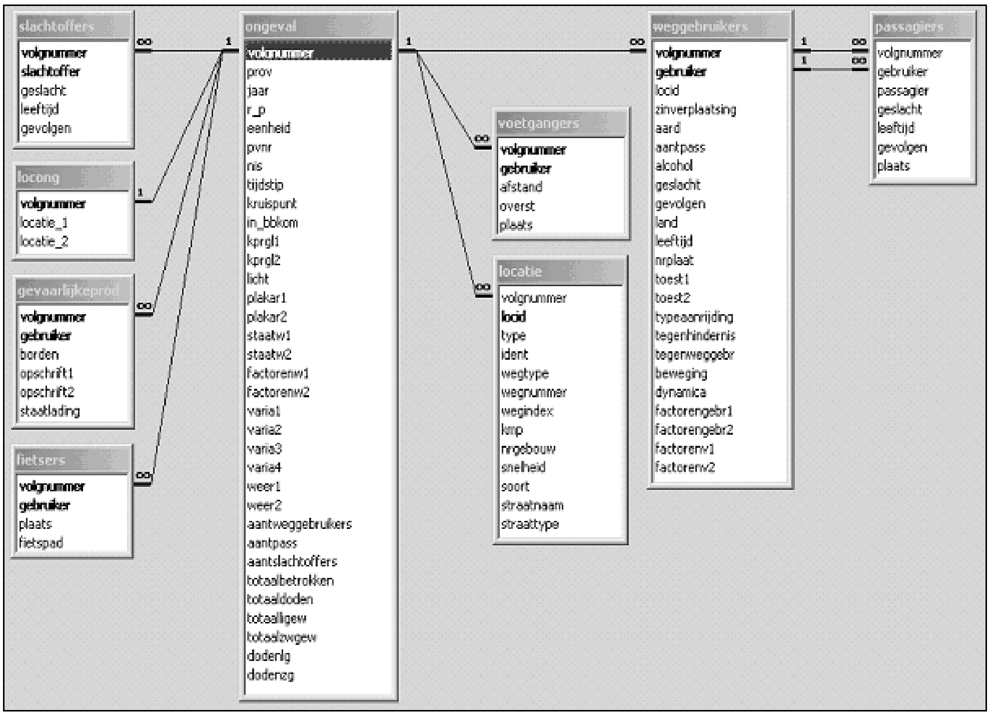

Figure 4.2: Structure of relational database on road accidents with casualties

that occurs on a numbered road can be linked to a hectometer or kilometer mark. An accident that occurs on a non-numbered road can be located by means of the street name and the house numbers. Again, this table can be related to the table ONGEVAL by means of the unique accident ID. Appendix 8 refers to the table LOCONG, which summarizes the unique accident location identification numbers for each accident. Finally, Appendix 9 shows the items of the table GEVAARLIJKE PROD, in which the information is listed about any dangerous products that were transported by road users who where involved in an accident with casualties on a public road. This information can be linked to the accident and to the road user by means of the unique accident ID and the road user number.

4.4.2 Additional Data

Additional traffic data are measured and collected by several research institutes in Belgium. For example, the Administration of Roads and Traffic (AWV) regularly measures the traffic intensities on district and province roads in Flanders. Furthermore, on these roads, AWV collects, in positive and negative direction, all visible characteristics that are related to infrastructure and the condition of the road surface (schools, bridges, sound walls, crah barriers, borders,...). However, the collection of these data is still in process and these characteristics are therefore not yet available for all district and province roads in Flanders. Furthermore, the research group Spatial Applications Division Leuven (SADL) collects and analyzes satellite images to find more detailed information on the circumstances in which an accident has taken place. Again, this project is still in process and these data were not available for this analysis.

Additionnally, in Belgium information on road accidents with no casualties is collected by the insurance companies by means of the official document for 'Accidents with material damage'. Due to privacy and commercial reasons this information is currently not made available for researchers outside the own company. However, the occurrence of these accidents is also an important issue in the traffic safety problem. For example, in comparison with accidents with casualties, material damage accidents could give information in the difference in severity of an accident. Other data that could give useful information on the occurrence of road accidents, such as official police reports and the official Belgian vehicle database, are not made available either due to the Belgian privacy law .

Finally, a number of possible interesting traffic data are not measured and collected in Belgium. For example, if we want to investigate the accident risk on zebra crossings, it would be interesting to know how many pedestrians actually use a zebra crossing to cross the street.

Efforts should be made to show the importance of additional data on road accidents such as data on accidents with material damage, official police reports, traffic intensities, etc. For example, international studies in which these data are used to enhance the traffic safety policy can be used to lay pressure on the proper authorities to make these data available for scientfic research. Note, however, that if, in future, different data sources will be used for analysis, special attention should be given when combining these data sets. These databases will not always be compatible in terms of format, period under study or research units.

4.4.3 Implications and Limitations

For the research in this chapter, we used the data that were available for analysis, in particular the data on accidents with casualties on a public road in Belgium (see Appendix 1 to Appendix 9). More specifically, in the following empirical studies, we use different subsets of this data set on road accidents with casualties to explore accident patterns, depending on the objective of the research.

An initial analysis indicated that these traffic accident data are highly skewed. This means that some of the attributes will have an almost constant value for each of the accidents in the database. However, as explained in the previous section, this will have no effect on the validity of the results since the association algorithm produces an interest measure that corrects the interestingness of each rule by taking the frequency of the attributes in the data set into account.

Furthermore, if we want to do research that leads to valid and reliable results, it is important that the data are reliable. The reliability of our data set will first of all be influenced by the objectivity of the police officer that fills out the accident form. Indeed, some variables that are listed in this document are rather subjective to decide on. Examples include the light conditions and the road surface characteristics. Additionally, the emotional atmosphere that usually coincides with the occurrence of an accident could influence the objectivity of the police officer. Therefore, we should question the reliability of this data before making any hasty conclusions. Secondly, persons who are involved in an accident could deliberately influence the reliability and validity of the reported accident data. For example, people will not always admit their faults and they will even be prepared to make a false statement about the accident circumstances for insurance reasons. Furthermore, the field ‘various’ referring to safety restraint use, cell phone use or other circumstances that could have played an important role in the accident or the related accident injuries is badly registered on the Belgian ‘Analysis Form for Traffic Accidents’. This type of error will give a distorted view of reality. Therefore, we should try to take these systematic errors into account when interpreting the results and be carefull in formulating conclusions concerning the real cause of the road accident .

Note however, that this poor quality of secondary road accident data is not just a problem in Belgium. Efforts are being taken on an international level to improve the quality of these data by constructing an international accident form (Workshop Mining Official Data, Finland, August 2002).

Finally, when interpreting and evaluating the results of our analyses, we should keep in mind that currently the total population of accidents in our research analyses equals the accidents with casualties. Therefore, the results of these analyses can not be generalized to all road accidents since these accidents also include accidents with only material damage. If we want to generalize our results to all accidents with casualties, we also have to be aware of the underregistration problem in traffic accident data. This is especially a problem for accidents with only lightly injured persons and small material damage. Road users will prefer to pay for the damage without any legal interaction. Therefore, the accidents that are reported to the police are just a selection of all accidents with casualties.

4.5 Profiling Accident Types

In this empirical study, the data mining technique of association rules is used to obtain a descriptive analysis of the accident data by identifying accident circumstances that frequently occur together. By analyzing the produced set of rules, describing underlying patterns in the data, we aim to discern frequently occurring accident types and identify relevant variables that make a strong contribution towards a better understanding of accident circumstances. Hereby, the emphasis will not only lie on the acquired interestingness of the generated patterns, but also on the interpretation of the results, which will be of high importance for improving traffic policies and ensuring traffic safety on the roads.

Analogously with the KDD process, we distinguish three steps in the mining process: a preprocessing step in which the available data are prepared for the optimal use of the mining technique, a mining step for generating the association rules and a post-processing step for identifying the most interesting association rules.

4.5.1 Preprocessing the Data Set

Although Belgium is a quite small country, densely populated and mainly urbanized (10.3 millions inhabitants; 30,528km^2), large disparities exist within the country (cf. Merenne, 1997). Hence, our analysis is limited to one administrative region of Belgium, i.e. the Brussels Capital region. This administrative region, consisting of 19 cities and municipalities among which the Belgian capital, covers an area of 161.4 km^2 and counts almost one million inhabitants. Furthermore, the Brussels Capital region covers 1,881 kilometers road which accounted for 3.18 billion vehicle kilometers traveled in 2002 (De Groote and Truwant, 2003).

In particular, this empirical study is based on the data set of traffic accidents obtained from the National Institute of Statistics (NIS) over a six year period (1991-1996) for the region of Brussels (Belgium). More specifically, for this region, we will concentrate on the accidents that can easily be located by the hectometer mark, i.e. the numbered roads or highways, national and provincial roads linking towns. Selecting these records from the data set resulted in a total of 10,672 traffic accident records.

Additionally, to explore the different accident types in this data set, we did not use all the available traffic accident records. Since our prime interest in this dissertation lies in the identification and understanding of hazardous locations, only the traffic accidents that occurred at a 'high frequency accident location' were selected for the analysis. Taking into account the average number of accidents per location for this data set, a criterion of minimum 10 accidents per location was used to identify these high frequency accident locations. This resulted in a total of 1,110 traffic accident records that were included in the analysis.

Furthermore, in the present data set, some attributes have a continuous character. Discretization of these continuous attributes is necessary, since generating association rules requires a data set for which all attributes are discrete. Therefore, the observations for these variables are divided into different intervals by grouping them into partitions. For example, four new attributes were created from the continuous variable time of accident': morning (6h-11h), afternoon (12h-16h), evening (17h-23h) and night (24h-5h). Another example is the discretization of the continuous variable 'maximum allowed speed'. The intervals for this variable were created on the basis of

our common knowledge of traffic speed regulations in Belgium: <50 km/hour, 50-65 km/hour, 70-90 km/hour, 100-120 km/hour. For those variables where no domain knowledge for grouping the attributes could be found, we used the Equal Frequency Binning discretization method to generate intervals containing an equal number of observations (Holte, 1993). Furthermore, attributes with nominal values had to be transformed into attributes with binary attribute values. This means that dummy variables had to be created by associating a binary attribute to each nominal attribute value of the original attributes. For example, for the variable 'intersection', the nominal attribute values 'near intersection' and 'outside intersection' were transformed into binary attribute values so that for each accident, one of these binary attribute values will receive the value '1' and the other the value '0'. Finally, irregularities such as data inconsistencies, missing values, redundant variables and double counts are tracked, listed and removed from the data set.
Examples of data inconsistencies are:

- The variable 'number of persons involved' = 1, the variable 'number of persons injured' = 40.
- The variable 'pedestrian' = false (there was no pedestrian involved in the accident), the variable 'type of collision' = collision with a pedestrian.

Examples of missing values are:

- The variables 'time of accident', 'location', are not filled out in the data set.
- The variable 'obstacle' = true (there was no obstacle on the location of the accident), the 'type of obstacle' = missing (no value was filled out).

Examples of redundant data are:

- The variable 'code unit': this variable indicates which unit of the police force has reported the accident. This kind of information will not be included as an exploratory or descriptive variable to model the occurrence of traffic accidents.

4.5.2 Generating Association Rules

A minimum support of 5 % was chosen for the analysis. This means that no item or set of items will be considered frequent if it does not appear in at least 56 traffic accidents. It could be argued that this choice for the support parameter is rather subjective. This is partially true, however a trial and error experiment indicated that setting the minimum support too low, leads to an exponential growth of the number of items in the frequent item sets. Accordingly, the number of rules that will be generated, will cause further research on these results to be impossible due to memory limitations. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable of generating trivial rules. For example, for a minimum support of 1 % the algorithm generated more than 2 million association rules. From this analysis, with a minsup = 5 % and minconf = 30 %, the algorithm obtained 101,861 frequent item sets of maximum size 4 (due to computational reasons) for which 313,663 association rules could be generated. These rules are further processed to select the most interesting rules.

4.5.3 Postprocessing the Association Rules Set

As explained in equation (4.3), the statistical rule significance (T) measures the validity of a rule, using the χ^2 test for statistical independence. This value can be negative (negative interdependence), neutral (independence) or positive (positive interdependence). Selecting the rules with a positive or a negative statistical significance from the association rules set narrowed down the results from 313,663 rules to 1,337 association rules.

These rules were further post-processed by ranking them on their Lift value and assessing the additional information of the rules towards this traffic accident analysis. An example of a rule that has no additional value is:

Wet⇒ Rain (sup = 19.91%, conf = 68.21%, T = +, L = 3.43).
This rule has an interest value of 3.43 and a positive rule significance indicating that whenever an accident happens, the probability of observing rain increases strongly if the road surface is wet. The confidence value shows that the probability of observing this kind of weather is 68.21% if the antecedent is true (wet), meaning that 68.21% of the days that an accident happens on a road with a wet surface it will be raining. The support of the rule indicates that 19.91% of all the accidents that occurred, happened on a wet road surface while it was raining. However, since it is obvious that even when no accidents happen the probability for rain will be higher with a wet road surface, this rule does not give an additional value towards a better understanding of the circumstances in which these accidents have happened.

4.5.4 Results

This section will give an overview of the most important results from the association analysis. More specifically, five topics highlighting different aspects of traffic accidents will be discussed: collision with a pedestrian, collision in parallel, sideways collision, week/weekend accidents and weather conditions. For each topic, the results will refer to the rule numbers [N] of the concerning rule table in which the rules are presented on the basis of a rank ordering of their lift value.

4.5.4.1 Collision with a Pedestrian

Table (4.4) illustrates that in 60.78% of all accidents involving pedestrians, collisions occur on intersections with traffic lights [7]. Moreover, accidents with pedestrians have a higher probability than expected of occurring at daylight [9], during the week [8] and in the afternoon [5].

Additionally, the results show that the pedestrian is often not coming unexpectedly from behind an obstacle through which he would not be visible at the moment of impact [1]. In 49.02%, he crosses the road on a zebra crossing with traffic lights for pedestrians [2] and his walking distance between sheltered places will more than expected lie between 6 and 12 meter [3]. This distance could relate to the length of the zebra crossing.

At first sight, these results may look surprising, since under these circumstances the pedestrian should be well visible for the other road users. Only in 13.07% of the accidents, the pedestrian will come from behind an obstacle through which he is not visible for the other road users at the moment of impact. Moreover, only 4.5% of the

Table 4.4: Rules for collision with a pedestrian

N	BODY		HEAD	T	L	sup	conf
1	pedestrian	⇒	visible	+	7.25	8.65	62.75
2	pedestrian	⇒	zebra crossing with traffic lights	+	7.25	6.76	49.02
3	pedestrian	⇒	unsheltered walking distance between 6 and 12 meter	+	7.25	5.59	40.52
4	pedestrian	⇒	1 road user upwards, 1 transverse	+	3.16	5.50	39.87
5	pedestrian	⇒	afternoon	+	1.29	5.50	39.87
6	pedestrian	⇒	1 roadway	+	1.26	10.99	79.74
7	pedestrian	⇒	intersection with traffic lights	+	1.24	8.38	60.78
8	pedestrian	⇒	week	+	1.20	11.53	83.66
9	pedestrian	⇒	daylight	+	1.19	10.00	72.55
10	pedestrian	⇒	driving in a straight direction	-	0.82	10.00	72.55
11	pedestrian	⇒	constant speed	-	0.75	8.11	58.82

collisions with a pedestrian occur while the pedestrian is crossing the street on a road with no zebra crossing, 13.72% while he is walking on a zebra crossing without traffic lights and 16.34% when he crosses the road walking next to a zebra crossing with traffic lights. A possible explanation for these results could be the large number of children that head for school, and therefore will be on the Belgian roads, around these times. These results can be linked with the results from Lee and Abdel-Aty (2005) who found that higher average traffic volume at intersections increases the number of pedestrian crashes.

Furthermore, the rules show that collisions with pedestrians mainly occur on roads with just one roadway [6] and one road user is more frequently than expected moving upwards in the street whereas the other road user is moving transversal on this direction [4]. The latter rule will probably relate to the walking direction of the pedestrian in relation to the moving direction of the driver since the Belgian Analysis Form for Traffic Accidents states that when the pedestrian is crossing the street while being involved in an accident, the pedestrian is moving in a transverse direction.

Finally, a collision with a pedestrian occurs less often than expected in the presence of a road user that drives at a constant speed [11] or when at least one vehicle is driving in a straight direction [10]. This arouses the suspicion that pedestrians will have a higher probability of getting hit by a vehicle when the road user is making a manoeuvre.

In conclusion, in the region of Brussels, collisions with pedestrians are a frequently occurring accident pattern. More specifically, these accidents will have a higher probability of occurring on crossroads with traffic lights, more specifically when the pedestrian is crossing the street on a zebra crossing with traffic lights, being well visible, at daylight, in the afternoon, during the week and when the road user is making a manoeuvre. These results can be linked with the results obtained by LaScala et al. (2000), who found that injuries in pedestrian-involved collisions are most likely to occur in areas of the city with a great population density. Ali (2001) adds that both pedestrians and drivers bear the responsibility equally for being involved in pedestrianvehicle crashes. More specifically, not paying attention is one of the most common causes among pedestrians while many drivers often do not respect the right-of-way of pedestrians.

4.5.4.2 Collision in Parallel

Table (4.5) shows that when an accident happens as a consequence of not respecting the distance between the different road users, the collision will almost inevitably take place between vehicles driving in the same direction [12]. From the definition of the Belgian Analysis Form for Traffic Accidents, this type of accident usually relates to a collision at the back of a vehicle but it can also be a collision between vehicles driving next to each other following the same direction. Additionally, the rule stated above is also valid in the opposite case [13] and in 42.36% of the collisions in parallel, one of the road users will have used his brakes with the intention to stop [14]. This type of accident will probably occur mostly in case of a collision at the back of a vehicle.

Furthermore, a collision in parallel will occur less frequently than expected when only two people are involved in the accident [18] and not respecting the distance between different road users will often lead to more than one collision [17]. Finally, a collision in parallel will less often than expected coincide with a road user driving at a constant speed [16] and will have a smaller probability than expected of happening at a crossroad [15].

To summarize, collisions in parallel will often be related with not respecting the distance between road users and with using the breaks with the intention to stop. However, this type of collision will have a smaller probability than expected of occurring on crossroads, at constant speed, with only two persons involved. To minimize the likelihood of being involved in this type of accident, Abdel-Aty and Abdelwahab (2004) suggest that a driver should maintain a space cushion that is appropriate for the driving conditions. A proper space cushion must provide a driver time to see and recognize a hazard and make a decision regarding what should be done. Then, there must be adequate space to bring the vehicle to a stop.

Table 4.5: Rules for collision in parallel

N	BODY		HEAD	T	L	sup	conf
12	distance	⇒	parallel	+	6.28	5.95	81.48
13	parallel	⇒	distance	+	6.28	5.95	45.83
14	parallel	⇒	brake	+	3.27	5.50	42.36
15	parallel	⇒	near intersection	-	0.95	11.98	92.36
16	parallel	⇒	constant speed	-	0.83	8.38	64.58
17	distance	⇒	1 collision	-	0.84	5.13	70.37
18	parallel	⇒	2 persons involved	-	0.81	8.56	65.97

4.5.4.3 Sideways Collision

The rules in table (4.6) indicate that when a sideways collision occurs, the road user will often not have respected the priority regulation of the crossroad [30]. Most of the times he will also drive at a constant speed [33]. These sideways collisions where the priority regulation of the intersection is not respected, have a higher probability than expected of happening on intersections where the road users should give way to the vehicles coming from the right [25]. When the priority on the intersection is regulated by traffic lights, the sideways collision will often occur when a road user makes a left

Table 4.6: Rules for sideways collision

N	BODY		HEAD	T	L	sup	conf
19	no priority + intersection with priority to the right	⇒	local road	+	3.02	10.54	43.82
20	no priority + intersection with traffic lights	⇒	left turn	+	2.48	14.00	75.24
21	intersection with traffic lights + road users opposite direction	⇒	left turn	+	2.12	7.84	64.44
22	no priority + intersection with priority to the right + local road	⇒	equal road functions	+	2.48	14.00	75.24
23	intersection with traffic lights + sideways collision	⇒	left turn	+	1.73	14.41	52.63
24	left turn	⇒	intersection with traffic lights	+	1.45	21.53	70.92
25	no priority + sideways collision	⇒	intersection with priority to the right	+	1.43	19.46	49.43
26	intersection with traffic lights + night with public lighting	⇒	sideways collision	+	1.34	5.77	80.00
27	intersection with priority to the right	⇒	no priority	+	1.34	24.05	69.35
28	left turn	⇒	no priority	+	1.29	20.36	67.06
29	no priority	⇒	sideways collision	+	1.27	39.37	75.87
30	sideways collision	⇒	no priority	+	1.27	39.37	66.01
31	1 road user upwards, 1 downwards	⇒	sideways collision	+	1.25	17.75	74.34
32	left turn	⇒	sideways collision	+	1.18	21.35	70.33
33	sideways collision	⇒	constant speed	+	1.09	50.63	84.89
34	1 road user upwards, 1 transverse	⇒	sideways collision	-	0.83	6.22	49.29
35	intersection with traffic lights + sideways collision	⇒	daylight	-	0.80	5.95	48.53
36	intersection with traffic lights	⇒	no priority	-	0.73	18.60	37.87
37	break	⇒	sideways collision	-	0.65	5.05	38.89

turn [23].

These results refer to the relation between not respecting the priority regulation of the intersection and the type of the priority regulation. An accident that occurs on an intersection where the road users should give way to the vehicles coming from the right, often coincides with a road user that does not respect this priority regulation [27]. An accident that occurs on an intersection with traffic lights will on the contrary less frequently coincide with not respecting this priority regulation [36]. In 75.24% of the accidents where this violation does occur with traffic lights, a road user will also have made a left turn [20]. Moreover, 70.92% of all accidents that occur when a road user makes a left turn, take place on an intersection with traffic lights [24].

Unfortunately, there is no information about which road user made the traffic violation, but it could be expected that the road user that turns left will not have respected the priority regulation. Additionally, not giving priority to the right has a higher probability than expected of occurring on intersections where at least one of the roads is local [19] or where both of the roads have a local character [22]. These results could indicate that the local character of a road could lead towards a misplaced feeling of traffic safety, whereas bigger, more important roads could enhance the concentration of the road users.

In general, 70.33% of the accidents where a road user turns left will lead to a sideways collision [32] and often when making a left turn, a priority violation will be the cause of the accident [28]. Not respecting the priority regulation of the intersection will lead in 75.87% of the accidents to a sideways collision (29).

Furthermore, when an accident occurs at night with public lighting and the road user approaches the traffic lights; the accident will often be a sideways collision [26]. This type of collision near the traffic lights will less frequently occur at daylight [35]. These results will probably relate to visibility that will be smaller at night.

A remarkable result is that when one road user is moving upwards in the street and another road user is moving in the opposite direction, the occurring accident will most of the times be a sideways collision [31]. We would rather expect that this road situation would lead to a frontal collision. However, a possible explanation could be that most drivers dry to avoid a head-on collision with a rapid correction, putting the vehicle into a sideways collision. Furthermore, these accidents will occur on intersections with traffic lights where the road users will drive in opposite directions and at least one of them will make a left turn [21].

Finally, when one of the road users uses his brakes with the intention to stop [37] or when one vehicle is driving upwards in the street and another vehicle is driving transversal on this direction [34] the accident will less frequently be a sideways collision.

In conclusion, there are two types of sideways collisions. The first type takes place at intersections where road users should give priority to the right. These accidents will most of the times be caused by not respecting this priority regulation. The second type of sideways collisions occurs on intersections with traffic lights. This type of accident will often be related with a road user making a left turn and will also frequently occur when the road users are moving in an opposite direction. These results can be linked with the results obtained by Larsen and Klines (2002) who showed that in left-turn collisions, some of the problems are related to unconscious, attention errors when approaching and entering an intersection.

4.5.4.4 Week/Weekend Accidents

Weekend: from Friday 21h- Monday 6h
Morning: 6h-11h; Afternoon: 12h-16h; Evening: 17h-23h; Night: 24h-5h

As shown in table (4.7), most accidents that occur at night, will take place during the weekend [38]. Moreover, the accidents that take place in the weekend will more often than expected occur at night with public lighting [39] and will less frequently occur at daylight [48]. Similarly, the accidents that happen on a Sunday will have a higher probability than expected of occurring at night with public lighting [40], in spite of the fact that on this day a lot of people will also be on the roads in the morning and in the afternoon, making so-called day trips or family excursions.

However, accidents that happen at night do less frequently than expected coincide with a driver whose physical condition is normal [47]. He will have a higher probability of being drunk, under the influence of drugs or just being exhausted or unwell.

In contrast, accidents that occur during the week with one road user driving upwards in the street and another road user driving in the transverse direction, usually take place at daylight [41]. In general, accidents that occur at daylight will most of the times take place during the week [44].

These results can be linked with the many studies that show that driving at night is more risky in terms of crash involvements per distance traveled than driving during the day (Keall et al., 2005). The reasons for this include the more prevalent use of alcohol by drivers at night, the effects of fatigue on the driving task and the risk

Table 4.7: Rules for week/weekend accidents

N	BODY		HEAD	T	L	sup	conf
38	night	⇒	weekend	+	1.92	7.38	58.57
39	weekend	⇒	night with public lighting	+	1.45	14.59	47.93
40	Sunday	⇒	night with public lighting	+	1.36	5.77	45.07
41	week + 1 road user upwards, 1 transverse	⇒	daylight	+	1.33	8.01	80.91
42	pedestrian	⇒	week	+	1.2	11.53	83.66
43	crossing important local road	⇒	week	+	1.14	11.17	78.98
44	daylight	⇒	week	+	1.11	46.76	76.89
45	afternoon	⇒	week	+	1.1	23.6	76.61
46	week	⇒	afternoon	+	1.11	23.6	33.94
47	night	⇒	normal physical condition	-	0.92	10.72	85.00
48	weekend	⇒	daylight	-	0.76	14.05	46.15

associated with reduced visibility.

As mentioned earlier, a collision with a pedestrian will also often occur during the week [42]. Even more, accidents that happen during the week will have a higher probability than expected of occurring in the afternoon [44]. The number of accidents that take place in the afternoon is accordingly smaller during the weekend than during the week [45,46].

Finally, 78.98% of the accidents that occur on crossroads where the crossing street is an important local road, take place during the week [43].

4.5.4.5 Weather conditions

Table (4.8) illustrates that accidents on a wet road surface will have a higher probability than expected of occurring at night with public lighting [50] and a smaller probability of occurring at daylight (58). Accordingly, accidents that happen at night with public lighting will coincide more frequently than expected with a wet road surface [51] and less frequently with a dry road surface [56].

Similarly, accidents that take place in the rain will have a higher probability of occurring at night with public lighting [49] and a smaller probability of occurring at daylight [59]. These results can be linked with the results of Andrey et al. (2001) who found that much of the elevated risk during rainfall appears to be related to visibility, since collision rates quickly return to near-normal after the rain has stopped, even if roads continue to be wet.

Furthermore, accidents that happen on a wet road surface [52], when it rains [54] or that occur at night with public lighting will less frequently coincide with a driver

Table 4.8: Rules weather conditions

N	BODY		HEAD	T	L	sup	conf
49	rain	⇒	night with public lighting	+	1.48	79.73	48.87
50	wet	⇒	night with public lighting	+	1.36	13.15	45.06
51	night with public lighting	⇒	wet	+	1.36	13.15	39.78
52	wet	⇒	normal physical condition	-	0.96	25.77	88.27
53	rain	⇒	no alcohol	-	0.96	18.56	93.21
54	rain	⇒	normal physical condition	-	0.95	17.39	87.33
55	night with public lighting	⇒	normal physical condition	-	0.94	28.65	86.65
56	night with public lighting	⇒	dry	-	0.84	19.64	59.4
57	>2 lightly injured persons	⇒	dry	-	0.82	5.13	57.58
58	wet	⇒	daylight	-	0.81	14.41	49.83
59	rain	⇒	daylight	-	0.74	8.92	44.8

whose physical condition is normal. Moreover, accidents in the rain have a smaller probability of occurring with a driver of whom the alcohol test will be negative or not required [53]. Finally, when more than two people are lightly injured, the accident will less frequently than expected have occurred on a dry road surface [57].

In conclusion, accidents that happen in the rain or on a wet surface will more frequently occur at night with public lighting. These accidents will also have a higher probability of occurring with a driver whose physical condition is not normal and a smaller probability of coinciding with a driver of whom the alcohol test will be negative or not required [53].

4.5.5 Conclusions

In this empirical study, the technique of association rules was used on a data set of traffic accidents for the region of Brussels for the period 1991-1996. The analysis showed that by generating association rules the identification of accident circumstances that frequently occur together is facilitated, leading to a strong contribution towards a better understanding of the occurrence of traffic accidents.

Furthermore, the results indicate that the use of the association algorithm allows to discern different accident types, each with different relevant accident conditions. In particular, zebra crossings with traffic lights and pedestrian visibility are important aspects of pedestrian collisions. Distance between the road users is an important aspect for collisions in parallel. Next, priority to the right and making a left turn are the most important factors in sideways collisions. Additionally, most accidents that occur at night, will take place during the weekend and finally, accidents that happen in the rain or on a wet surface will more frequently occur at night with public lighting.

4.6 Profiling High Frequency Accident Locations

In this section, the technique of association rules is used to perform a comparative analysis between high frequency and low frequency accident locations. The objective of this empirical study is to determine the discriminating character of the accident characteristics of dangerous accident locations.

4.6.1 Preprocessing the Data Set

This study is based on the large data set of traffic accidents obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the year 1999. As explained earlier, these data are obtained from the Belgian 'Analysis Form for Traffic Accidents' that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. Given the limitation that this one year period is not long enough to limit random fluctuations, the period under study does limit changes in road and traffic conditions. In total, for this period 34,353 traffic accident records are available for analysis. On average, 45 attributes are available for each accident in the data set.

To discern high frequency accident locations from low frequency accident locations, each accident needs to be linked with a location parameter that corresponds to a unique geographical location. Therefore, the accidents that occurred at a highway or a district or province road are located by means of the road identification number and the hectometer mark. The accidents that took place at a non-numbered road are located using the street name and the name of the city in which the accident occurred.

Next, two different data sets were selected to explore association relationships between traffic accident attributes. Since our prime interest lies in the profiling and understanding of dangerous accident locations, only the traffic accidents that occurred at a high frequency accident location were selected for the first analysis. This allows us to give a descriptive analysis of frequently occurring accident patterns on highly concentrated accident locations. To identify these locations, a criterion of minimum five accidents per location was used. This resulted in a total of 3,368 traffic accident records that were included in the first analysis. This number of accidents corresponds with the fact that in Flanders 15% of all the traffic accidents occur on so-called 'dangerous spots' (Ministry of the Flemish Community, 2001). The second association analysis is carried out on the remaining low frequency accident locations, including 30,985 accidents. By comparing the results from these two analyses, we can determine the discriminating character of the accident characteristics of high frequency accident locations.

As explained in the previous empirical study, discretization of the continuous attributes is necessary, since generating association rules requires a data set for which all attributes are discrete. Therefore, the observations for these variables are divided into different intervals by grouping them into partitions. on the basis of expert knowledge. For those variables where no domain knowledge for grouping the attributes could be found, we used the Equal Frequency Binning discretization method to generate intervals containing an equal number of observations (Holte, 1993). Furthermore, attributes with nominal values had to be transformed into attributes with binary attribute values. Finally, irregularities are removed from the data sets.

4.6.2 Generating Association Rules

A minimum support value of 5 % was chosen for the analysis. This means that no item or set of items will be considered frequent for the first analysis if it does not appear in at least 165 traffic accidents. Obviously the same threshold will be used for the second analysis since the main purpose of this research involves a comparative analysis between the rules sets of the high frequency accident locations and the low frequency accident locations. Again, it could be argued that the choices for the values of these parameters are rather subjective. As explained in the previous study, this is partially true, however a trial and error experiment indicated that setting the minimum support too low, leads to exponential growth of the number of items in the frequent item sets. Accordingly, the number of rules that will be generated will cause further research on these results to be impossible due to computer memory limitations. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable of generating trivial rules. Analogously, the minimum confidence value was set at 30 %. This means that a rule is considered reliable when the consequent of the rule occurs at least one out of three times that the antecedent appears. By choosing different confidence values, a trial and error experiment showed that this parameter value gives rather stable results concerning the amount of rules generated by the algorithm.

From the high frequency accident locations, with a minsup = 5 % and minconf = 30 %, the algorithm obtained 187,829 frequent item sets of maximum size 4 (due to computational reasons) for which 598,584 association rules could be generated. Although these results relate to a relatively small number of accident records, they are quite reasonable since an average of 40 items is available per accident, allowing the algorithm to generate multiple combinations of size 4 item sets. With the same parameters the second analysis resulted for the low frequency accident locations in 183,730 frequent item sets of maximum size 4 for which 575,974 association rules could be generated. These rules are further processed to select the most interesting rules.

4.6.3 Postprocessing the Association Rules Set

As explained earlier, the purpose of postprocessing the association rules set is to identify the subset of interesting (i.e., non-trivial) rules in a generated set of association rules. Selecting the rules with a positive or a negative statistical significance (see equation 4.3) from the association rules set narrowed down the results to 14,690 association rules for the high frequency accident locations and 77,282 association rules for the low frequency accident locations.

However, after ranking the association rules on their lift value and removing the non-significant rules from the rules set, one important problem still remains. The discovered accident patterns for the high frequency accident locations will give a description of the frequently occurring accident circumstances, but they may also be characteristic for the accidents that occur at low frequency accident locations as they represent the necessary but not the sufficient condition for the membership of a high frequency accident location. Therefore, we use the interestingness measure to limit the association rules to only the discriminating or useful ones (see equation 4.2). This interestingness measure is based on the deviation of the characteristic rules discovered for the accidents that occurred on high frequency accident locations (with support s_h) from the accidents that occurred on low frequency accident locations (with support s_l). Since in this research we are mainly interested in profiling the high frequency accident locations, we will pay special attention to the rules with a positive interest value, i.e. approximating '1'.

4.6.4 Results

As stated earlier, the emphasis in this study lies on the identification and profiling of frequently occurring accident patterns at high frequency accident locations and the degree in which these accident characteristics are discriminating between high frequency and low frequency accident locations. Selecting the association rules that appear in both the high frequency accident rules set and the low frequency accident

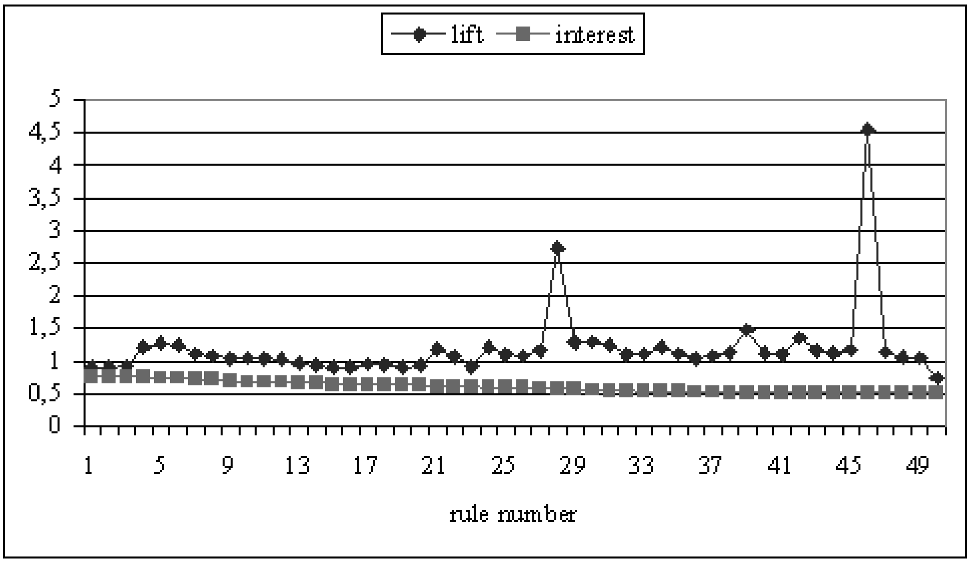

Figure 4.3: Association rules ranked on descending interest values

rules set results in 3,670 statistically significant association rules. These can be further post-processed by means of the interestingness measure. When ranking the association rules on their interest value, figure (4.3) shows, for the 50 most discriminating rules, that a high interestingness value does not inseparably correlates with a strong lift value. Note that we do not use the term 'high' for the lift value, since a very small lift value, i.e. considerably differing from 1, also indicates a strong (negative) dependency between the rule body and the rule head. These results show that accident characteristics that have the most discriminating power to identify high frequency accident locations are not necessarily the most interesting rules according to their lift values.

Accordingly, when ranking the association rules on the lift value, figure (4.4) and figure (4.5) indicate that although the association rules identify frequently occurring patterns that are descriptive for the occurrence of accidents on high frequency accident locations, they are not necessarily discriminating between the profile of high frequency accident locations and low frequency accident locations.

When looking at the association rules with the highest lift values, figure (4.4) shows that the majority of these rules have a small interestingness value. Also for the very small lift values, figure (4.5) indicates that the strong dependencies between

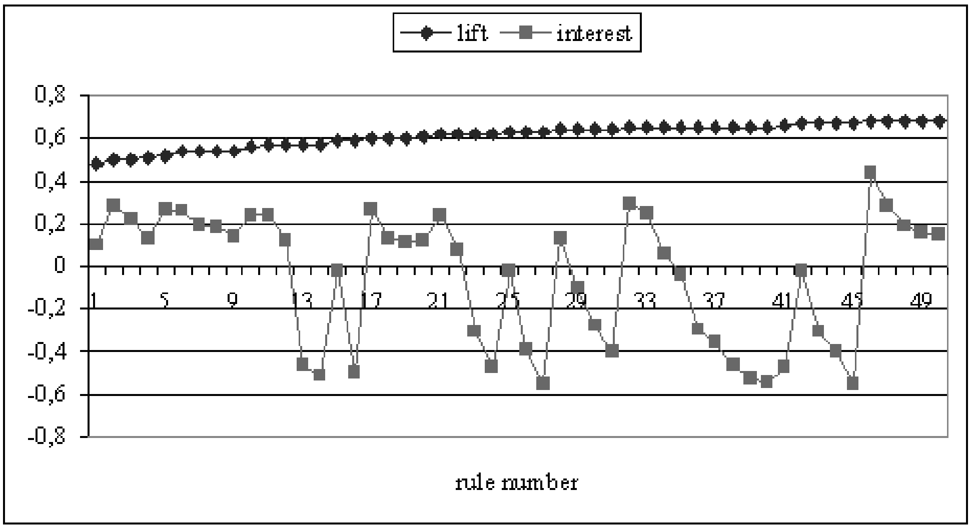

Figure 4.4: Association rules ranked on descending lift values

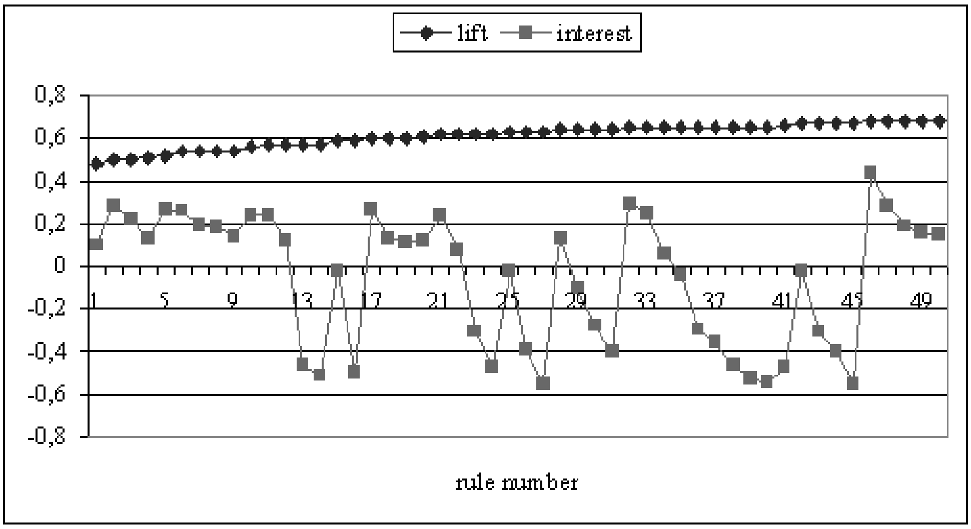

Figure 4.5: Association rules ranked on ascending lift values

different accident characteristics do not always correspond with high interestingness values. In general, these rules with a strong lift value and a low interestingness value are characteristic for both high frequency accident locations and low frequency accident locations.

When looking at the interpretation of the association rules, table (4.9) shows that the rules with the 10 highest interestingness values mostly relate to location characteristics of the accidents. This means that the most discriminating characteristics between high frequency accident locations and low frequency accident locations are related to infrastructure or location features. For example, when an accident occurs on a roadway with separated lanes (by means of a guard-rail or a roadside), it will less frequently than expected take place outside a crossroad [1-3]. Additionally, an accident occurring on this type of roadway will have a higher probability than expected of occurring outside the inner city [4-6]. Since the accidents analyzed in this research all occurred in the region of Flanders, these results seems quite reasonable since in this part of Belgium most roadways with separated lanes are located outside the inner city. In total 46.4 % of all accidents that occur at high frequency accident

Table 4.9: Rules with highest interest value for high and low frequency accident locations

N	BODY		HEAD	Int	T	L	s_h	s_l	conf
1	Belgian road user + roadway separated lanes	⇒	outside crossroad	0.76	-	0.90	36.46	8.51	70.62
2	road user normal condition + roadway separated lanes	⇒	outside crossroad	0.76	-	0.91	37.32	8.77	70.93
3	roadway separated lanes	⇒	outside crossroad	0.76	-	0.92	39.81	9.54	71.60
4	daylight + roadway separated lanes	⇒	outside inner city	0.75	+	1.21	30.87	7.81	81.91
5	Belgian road user + roadway separated lanes	⇒	outside inner city	0.74	+	1.27	44.12	11.40	85.45
6	roadway separated lanes	⇒	outside inner city	0.73	+	1.24	46.40	12.34	83.45
7	road user normal condition + roadway separated lanes	⇒	age 30-45	0.72	+	1.11	31.79	8.91	60.58
8	roadway separated lanes	⇒	age 30-45	0.72	+	1.08	32.98	9.31	59.32
9	roadway separated lanes	⇒	passenger car	0.69	+	1.04	47.86	15.02	86.07
10	road user normal condition + roadway separated lanes	⇒	Belgian road user	0.68	+	1.05	49.02	12.52	93.38

locations take place on a roadway with separated lanes outside the inner city [6]. In comparison with the low frequency accident locations, where only 12.34 % of the accidents can be attributed to this sort of location, these roadways with separated lanes outside the inner city are very characteristic for the occurrence of black spots and black zones. Although these results seem quite reasonable, they do indicate that roadways with separated lanes outside the inner city are an important problem for traffic safety. Therefore, further research on the cause of the unsafe character of these roads will be necessary. A high traffic intensity would seem the logical explanation for the high number of accidents that occur on these roads, but another possible explanation could also be the infrastructure of these roads. Indeed, several authors have demonstrated that dangerous accident sites often correspond to places where improvements could be made in terms of road geometry (see e.g. Agent and Deen, 1975; Wong and Nicholson, 1992; Taber, 1998; Martin, 2002; Greibe, 2003). Depending on the results, government could consider restructuring these roads or changing their traffic regulation.

Furthermore, when an accident takes place on a roadway with separated lanes, the road user will more frequently than expected be of the age 30 until 45 [7,8]. Finally, a characteristic pattern for high frequency accident locations is the involvement of at least one passenger car among the road users when the accident occurs on a roadway with separated lanes [9]. Accordingly to the previous results, this kind of accident accounts for 47.86 % of all accidents on high frequency accident locations, whereas the same accident circumstances only occur in 15.02 % of the low frequency accident locations.

Table (4.10) gives the 10 most important rules for the high frequency accident locations based on the descending lift values. The negative interest values of these rules indicate that the accident circumstances described in these patterns have a higher occurrence on the low frequency accident locations than on the high frequency accident locations, although the differences in support values are very small. Furthermore, the results show that the strongest positive dependencies between accident characteristics are not as strongly location related. Most of these association rules refer to the number of persons involved in the accident and the number of casualties following the accident [11-15,18-20]. This can be explained by the very small interestingness values of the rules, referring to the occurrence of the accident characteristics at both low frequency accident locations and high frequency accident locations. As a result, these rules will identify patterns that will be less geographical or location related

Table 4.10: Rules with highest lift value for high and low frequency accident locations

N	BODY		HEAD	Int	T	L	s_h	s_l	conf
11	0 passengers + 1 seriously injured	⇒	0 lightly injured	-0.31	+	8.28	7.33	10.56	85.76
12	0 deadly injured + 0 passengers + 0 lightly injured	⇒	1 seriously injured	-0.31	+	8.05	7.21	10.49	96.05
13	0 deadly injured + 0 deaths after serious injury + 0 seriously injured	⇒	1 seriously injured	-0.29	+	7.36	7.69	10.79	87.80
14	0 deadly injured + 0 lightly injured	⇒	1 seriously injured	-0.29	+	7.31	7.69	10.80	87.21
15	0 passengers + 0 lightly injured	⇒	1 seriously injured	-0.31	+	7.14	7.33	10.56	85.17
16	marked cycle track	⇒	one-way cycle track	-0.08	+	7.00	5.78	6.31	82.98
17	one-way cycle track	⇒	marked cycle track	-0.08	+	7.00	5.78	6.31	48.87
18	1 seriously injured	⇒	0 lightly injured	-0.28	+	6.41	7.92	11.02	66.42
19	0 lightly injured	⇒	1 seriously injured	-0.28	+	6.41	7.92	11.02	76.50
20	outside inner city + 1 seriously injured	⇒	0 lightly injured	-0.10	+	6.06	5.81	6.49	62.82

and more human or vehicle related. Two exceptions can be made. The association rules concerning cyclists do refer to frequently occurring infrastructure characteristics. These patterns state that when an accident occurs with a cyclist who is riding on a cycle track that is marked on the roadway, the type of this cycle track will more frequently than expected be a one-way track and vice versa [16,17]. This kind of accident accounts for 7% of all accidents on high frequency accident locations but also for 5.78% of all accidents on low frequency accident locations. Therefore, these accident characteristics are not very discriminating between high and low frequency accident locations, but they do however identify an important problem in the traffic safety of cyclists in general.

The results for the 10 strongest negative dependency rules are shown, based on the ascending lift values, in the lower part of table (4.11). The association rule with the smallest lift value indicates that when an accident occurs on a wet road surface, the weather will less frequently than expected be normal [21]. Obviously this is a very strong dependency that is valid for most accidents, whether they occur on high frequency accident locations or not. The second most important rule of this table indicates that when a road user crashes into an obstacle outside the roadway, he was less frequently than expected continuing his driving direction [22]. Furthermore, when losing control over the steering wheel, there is a smaller probability than expected

Table 4.11: Rules with highest lift value for high and low frequency accident locations

N	BODY		HEAD	Int	T	L	s_h	s_l	conf
21	wet road surface	⇒	normal weather	0.10	-	0.48	11.19	10.09	38.47
22	collision obstacle outside roadway	⇒	continuing driving direction	0.28	-	0.50	7.77	5.56	38.87
23	loss control steering wheel	⇒	2 road users	0.22	-	0.50	7.92	6.17	32.40
24	Belgian road user + 2 lightly injured	⇒	0 passengers	0.13	-	0.51	5.84	5.08	36.21
25	Belgian road user + loss control steering wheel	⇒	2 road users	0.26	-	0.52	7.42	5.47	34.01
26	passenger car + loss control steering wheel	⇒	2 road users	0.26	-	0.54	6.91	5.10	35.30
27	male road user + loss control steering wheel	⇒	2 road users	0.19	-	0.54	6.97	5.63	34.92
28	no alcohol + loss control steering wheel	⇒	2 road users	0.19	-	0.54	6.35	5.17	34.97
29	Belgian road user + 0 seriously injured + 2 lightly injured	⇒	0 passengers	0.14	-	0.54	5.84	5.03	38.33
30	road user normal condition + loss control steering wheel	⇒	2 road users	0.24	-	0.56	7.74	5.89	36.20

that two road users be involved in the accident [23,25-28,30]. Corresponding with the results for the positive rule dependencies, these patterns relate mainly to human characteristics and less to location related circumstances. However, a small difference in interpretation can be noted since the negative dependency rules mainly describe behavioral aspects of traffic accidents indicating different accident types. It seems that losing control over the steering wheel and crashing into an obstacle outside the roadway are frequently occurring accident patterns, each with different accident circumstances. Indeed, run-off-roadway accidents have already often been studied (see e.g. Lee and Mannering) and are often related to an inadequacy of the speed and/or behavior of the user to the driving circumstances. Accordingly, these association rules are more descriptive for the occurrence of accidents in general and they are less discriminating between high frequency accident locations and low frequency accident locations.

When looking at the results of the tables (4.9),(4.10) and (4.11) a final remark concerning the interest value of the rules should be made. One can see that the support values for the patterns in the low accident locations remain quite stable over the three tables. However, the support values for the rules concerning the high frequency accident locations are considerably high for table (4.9) and considerably smaller for the tables (4.10) and (4.11). Therefore, the increase in the interest value of the rules and consequently the increase in the discriminating character of the rules are mainly related to the strong occurrence of the accident circumstances in the high frequency accident locations and not as much by the weak occurrence of these patterns in the low frequency accident locations.

Special attention should also be given to the association rules that do appear in the rules set of the high frequency accident locations but not of the low frequency accident locations. Table (4.12) shows the results for these association rules with the 10 highest lift values. The rules are mainly related with the number of casualties of the accident [31-34] but also with one specific accident type: collision with an obstacle

Table 4.12: Rules for high frequency accident locations

N	BODY		HEAD	T	L	s_h	conf
31	0 deaths after serious injury + 0 passengers + 0 lightly injured	⇒	1 seriously injured	+	7.19	7.33	85.76
32	normal weather + 1 seriously injured	⇒	0 lightly injured	+	6.73	6.65	69.78
33	normal physical condition + 0 deaths after serious injuries + 1 seriously injured	⇒	1 seriously injured	+	6.71	7.30	80.13
34	normal physical condition + 0 lightly injured	⇒	1 seriously injured	+	6.67	7.30	79.61
35	roadway separated lanes + 0 seriously injured + collision obstacle outside roadway	⇒	crash barrier	+	6.51	5.52	52.39
36	roadway separated lanes + outside crossroad + collision obstacle outside roadway	⇒	crash barrier	+	6.34	6.74	51.01
37	roadway separated lanes + outside inner city + collision obstacle outside roadway	⇒	crash barrier	+	6.31	6.77	50.78
38	roadway separated lanes + 0 deaths + collision obstacle outside roadway	⇒	crash barrier	+	6.21	6.68	50.00
39	outside inner city + 0 seriously injured + collision obstacle outside roadway	⇒	crash barrier	+	6.05	6.56	48.68
40	roadway separated lanes + collision obstacle outside roadway	⇒	crash barrier	+	6.03	6.80	48.52

outside the roadway. When this kind of collision occurs on a roadway with separated lanes outside a crossroad or outside the inner city, the obstacle will more frequently than expected be a crash barrier (iron or concrete) [35-40].

In the most favorable situation, these rules are not characteristic for the low frequency accident locations (s_l=0, Interest=1) and consequently these patterns optimally discriminate between high frequency accident locations and low frequency accident locations. However, in the most pessimistic situation these patterns do exist for the low frequency accident locations but they do not appear in the association rules set due to the value of the minimum support parameter (5%). Moreover, the choice of the minimum confidence parameter (30%) could inhibit the algorithm from generating a number of rules although the frequent item sets exceed the minimum support parameter. Consequently, although these association rules could give some valuable information on the occurrence of traffic accidents in general, no conclusion can be made towards the discriminating character of these rules between high an low frequency accident locations. Furthermore, the support value (or the occurrence) of these accident circumstances is relatively low for the high frequency accident locations and therefore we will no further go in to detail about the interpretation of these rules.

4.6.5 Conclusions

In this study, the technique of association rules was used on a large data set of traffic accidents for the region of Flanders for the year 1999. The objective of this empirical study was to identify and profile frequently occurring accident patterns at high frequency accident locations and determine the degree in which these accident characteristics are discriminating between high frequency and low frequency accident locations.

First of all, we found that the most discriminating characteristics between high frequency accident locations and low frequency accident locations are related to infrastructure or location related circumstances. More specifically, results indicate that roadways with separated lanes outside the inner city are an important problem for traffic safety. However, further research on the cause of the unsafe character of these roads will be necessary. Next, results showed that the accident characteristics that occur frequently both at low frequency accident locations as at high frequency accident locations are less geographical or location related and more human or vehicle related indicating that human and behavioral characteristics play an important role in the occurrence of all traffic accidents. In particular, most of these association rules refer to the number of persons involved in the accident and the number of casualties following the accident. Furthermore, although these accident characteristics are not very discriminating between high and low frequency accident locations, we identified that losing control over the steering wheel, crashing into an obstacle outside the roadway and cyclists are important issues in the traffic safety problem in general.

In conclusion, this analysis shows that a special traffic policy towards high frequency accident locations should be considered, since these accident locations are characterized by specific accident circumstances, which require different measures to improve the traffic safety.

4.7 Profiling Black Zones

As explained in the introduction of this dissertation, methods developed for identifying accidents concentrations often apply to hot spots (also called black spots, hazardous locations, sites with promise etc.) which are pinpoint concentrations of road accidents that often migrate over time (see e.g. Silcock and Smyth, 1985; Maher, 1990; Nguyen, 1991; Joly et al., 1992; Hauer, 1996; Thomas, 1996 or Vandersmissen et al., 1996).

More recently, the identification of ‘black zones’ or hazardous road segments has been reconsidered in the literature (see Flahaut et al., 2003 for a review); they arise from the awareness of the spatial interaction existing between contiguous accident pinpoint locations. The existence of such road sections on which the number of accidents is high reveals spatial concentrations and hence suggests spatial dependence between individual accidents’ occurrences. In fact, these studies focus on a well-known exploratory spatial data analysis problem: the definition and the explanation of hot spots (see e.g. Levine, 2002 or Vistisen, 2002).

In this research, the location and the length of the black zones are defined by means of local spatial autocorrelation indices and they are considered as given in our problem. Therefore, the problem tackled here is not the definition of the black zone, but its exploration, i.e. to understand how road accidents cluster in hazardous road segments. More specifically, we are interested in finding out which factors are associated to the accidents in black zones by generating frequent item sets. In particular, accident circumstances that frequently occur together inside black zones will be identified. Furthermore, these patterns are compared with accident characteristics occurring outside those black zones. This allows the investigation of the differences between accident patterns inside and outside black zones, and hence to understand why spatial concentrations are observed. This way, we expose a number of hypotheses, which we then try to explain using other research studies and domain knowledge.

4.7.1 Preprocessing The Data Set

4.7.1.1 Description of the Data

Here, our analysis is here limited to one administrative region: the Walloon Brabant, which is a province extending South of Brussels. It is 1,100 km^2 large and counts almost 350,000 inhabitants and 4,604 kilometers of numbered roads. It is mainly characterized by urban sprawl, but also by the existence of some former small market towns like Nivelles, Braine-l'Alleud, Wavre or Jodoigne (Dekeersmaecker et al, 2004; Thomas et al., 2004). The Eastern part is still rural, the Western part more industrial. Limiting the extent of the studied area enables one to better control for other sources of variations (mobility habits, friction of distance, mobility policies, etc).

Analogously with the previous study, for this region, traffic accidents data are obtained from the 'Belgian Analysis Form for Traffic Accidents' for each road accident with casualties on numbered roads. Furthermore, the period under study is 1997-1999: it is long enough to limit random fluctuations in the accident counts and short enough to limit changes in road and traffic conditions. Several other researchers also used the same period of crash history records to address different road safety issues (Cheng and Washington, 2005).

These data indicate that for the region of Walloon Brabant, 1,861 injury accidents occurred between 1997 en 1999. In these accidents, 81 persons were deadly injured, 333 seriously and 2,374 lightly injured.

4.7.1.2 Definition of the Black Zones

The location and the extension of hazardous locations are here defined in a preceding study (Flahaut et al., 2003). Simply said, the hectometer (100 meters) is the smallest spatial unit for which road accident data are spatially available (stone markers on the numbered roads). This unit is very small and a source of many location errors; moreover, at this level of aggregation, black spots only refer to less than 10% of the total number of accidents and they are known to migrate over time.

We want to know if accidents are concentrated in space, if hectometers with large accidents records are scattered or clustered together. We therefore used the concept of local spatial autocorrelation. We know that spatial independence is an arrangement of accidents such that there are no spatial relationships between them. The intuitive concept is that the location of an accident is unrelated to the location of any other accident. The opposite condition - spatial autocorrelation - is an arrangement of accidents where the locations of the hectometers are related to each other, that is they are not statistically independent of one another. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated (Levine, 2002). When accidents are clustered together, we refer to this arrangement as positive spatial autocorrelation. Conversely, an arrangement where accidents are dispersed is referred to as negative spatial autocorrelation. Global autocorrelation gives then a rough idea of the general spatial arrangements of accidents.

The Moran's I statistic is one of the oldest indicators of spatial autocorrelation (Moran, 1948). We here used a local index developed as a local indicator of spatial autocorrelation (LISA) by Anselin in 1995. It takes high values for hectometers located close to each other and having large numbers of accidents. Closeness is here measured in terms of distance measured on the road network. Sensitivity analyses were performed to the way distance is measured (Flahaut and Thomas, 2002); the technique of local Moran I has also been compared to kernel methods for defining road accidents black zones (Flahaut et al., 2003). Moreover, stability of the spatial structure put forward with Local Moran I has been analyzed over time and space in the same studied area: the locations of the black zones remain comparable from one year to the other (Eckhardt, Flahaut and Thomas, 2004). These sensitivity analyses confirm the performance of the technique as well as a strong spatial structure of the road accidents in the Brabant Wallon area. This means that the location of the road accidents concentrations is not or only slightly dependent upon the method used and the period of time chosen. The spatial structure is strong; the methodological choices made for defining the hazardous locations should not affect the conclusions of this research. We refer to former studies for a critical analysis of the method (Flahaut, 2004a and 2004b; Flahaut et al., 2003; Flahaut and Thomas, 2002).

From a former analysis on the same studied area (Eckhardt, Flahaut and Thomas, 2004), we know that 47% of the road hectometers did never register any accidents. Furthermore, black zones represent 38% of the total number of accidents, but only 12% of the total number of hectometers. Between 1997 and 1999, 476 kilometers of black zones have been defined by local autocorrelation indices in the Walloon Brabant. Selecting the accidents that occurred inside the black zones results in a total of 553 road accidents. The second data set, containing the accidents that took place outside a black zone involves 1,287 road accidents. (Note: the belonging or not of an accident location to a black zone could not be defined for 21 accidents).

4.7.1.3 Transforming the Data Set

Two data sets are defined according to whether an accident belongs to a black zone or not. For each accident, the official form provides several variables related to the accident, the road-users and the place of the accident. Some variables have a continuous character. As explained in the previous sections, discretization of these variables is necessary, since generating frequent item sets requires a data set for which all items are discrete. Furthermore, attributes with nominal values have been transformed into binary attribute values and irregularities such as data inconsistencies, missing values, redundant variables and double counts are tracked, listed and removed from the data sets (Casaer et al., 2003).

In total, 292 items (characteristics of the accidents) are included in the analysis. These items give information on 45 variables of the accidents (see appendix 10). Since not every one of these variables is filled out for each accident, on average 40 of the 292 accident characteristics are available per accident.

4.7.2 Mining for Frequent Item Sets

A minimum support value of 5% was chosen for determining the frequent item sets. This means that no item or set of items will be considered frequent for the analysis of black zones if it does not appear in at least 27 road accidents. Obviously the same threshold will be used for analyzing the non-black zones since the main purpose of this research involves a comparative analysis between the accident patterns characterizing black zones and non-black zones. This means that an accident characteristic must appear in at least 64 accidents to be considered as frequent for the non-black zones. It could be argued that the choice of the values of these parameters is rather subjective. This is partially true, however a trial and error experiment indicated that setting the minimum support too low, leads to exponential growth of the number of items in the frequent item sets. In contrast, by choosing a support parameter that is too high, the algorithm will only generate trivial accident patterns.

From the data set containing the accidents that occurred inside a black zone, with a minsup of 5%, the algorithm obtained 187 761 frequent item sets of maximum size 4 (due to computational reasons). Although these results relate to a relatively small number of accident records, they are quite reasonable since an average of 40 items is available per accident, allowing the algorithm to generate multiple combinations of size 4 item sets. With the same parameters the analysis of the accidents that took place outside a black zone resulted in 181,066 frequent item sets of maximum size 4.

4.7.3 Postprocessing the Frequent Item Sets

As stated in the introduction of this empirical study, the emphasis in this study lies on the profiling of hazardous road segments in terms of accident related data and the degree in which these accident characteristics are discriminating between black zones and non-black zones. Therefore, we will first discuss the item sets that are unique for the accidents that occurred inside a black zone. Selecting these item sets resulted in 40,731 frequent accident patterns. These item sets represent very characteristic combinations of accident circumstances for hazardous road segments and can be considered as very discriminating between black zones and non-black zones. Next, we will discuss the items sets that are unique for the accidents that took place outside the black zones. In total, these correspond with 34,036 frequent item sets. Again, these unique accident patterns are good discriminators between black zones and non-black zones.

When discussing these accident patterns, we are mainly interested in the frequent item sets with lift values strongly differing from 1 since these item sets represent strong dependencies between the different items of the item set. However, note that we should not compare the absolute lift values of the item sets of different sizes, since the more items the item set consists of, the higher the lift value will become (see equation (4.1)). Accordingly, we will use different cut-off values for the lift parameter to determine the most interesting rules.

Finally, we will discuss the item sets that are frequent for both groups of accidents (in and out black zones). Selecting these frequent item sets resulted in 147,030 accident patterns. To determine the discriminating character of these accident patterns, these can be further post-processed by means of the interestingness measure.

4.7.4 Results

4.7.4.1 Accident Patterns inside Black Zones

Selecting the frequent item sets that are unique for accidents occurring inside a black zone and with very strong lift values results in 50 item sets of size 2 (lift <0.5 or lift

>5), 108 item sets of size 3 (lift <0.5 or lift >5) and 240 item sets of size 4 (lift <0.5 or lift >15). Table (4.13) gives an overview of the most interesting of these frequent item sets. In the remainder of this research, we will refer to the number of these item sets [N] when discussing the results.

A first result shows that intersections covered by traffic lights are often associated with hazardous road segments [see N=1 in table (4.13)]. More specifically, accidents on intersections covered by traffic lights frequently coincide with a road user making a left turn [2-5]. Accordingly, side impact collisions after making a left turn and giving no priority is a frequently occurring type of collision between two road users [6]. Note that this type of accident really characterizes black zones and absolutely

Table 4.13: Frequent item sets for accidents inside black zones

N	Item 1	Item 2	Item 3	Item 4	sup	L
1	near intersection	intersection traffic lights			9.4%	4.09
2	left turn	intersection traffic lights			5.4%	3.54
3	near intersection	intersection traffic lights	left turn		5.4%	14.51
4	district or province road	intersection traffic lights	left turn		5.2%	5.87
5	straight direction	intersection traffic lights	left turn		5.1%	5.17
6	2 road users	no priority	sideways collision	left turn	5.9%	15.36
7	rain	aqua planning			5.1%	3.42
8	wet road surface	rain	aqua planning		5.1%	7.88
9	built up area	pedestrian collision			5.6%	2.74
10	1 road lane	speed 50 KMH	built up area	pedestrian collision	5.6%	16.80
11	built up area	age road user 0_17			5.4%	2.73
12	1 road user	loss control to the left	against crash barrier		5.2%	7.96
13	1 road user	collision obstacle outside roadway	loss control to the left	against crash barrier	5.2%	26.37
14	speed 120 KMH	loss control to the left	against crash barrier		5.1%	6.53
15	speed 120 KMH	collision obstacle outside roadway	loss control to the left	against crash barrier	5.1%	21.65
16	road user normal condition	car	speed 50 KMH	loss control steering wheel	5.2%	0.49

not other locations. Moreover, accidents with left-turning vehicles can be divided into quite different accident situations. Consequently, the accident attribute 'left turn' will cover various problems. As explained before, these results can be linked with the results obtained by Larsen and Klines (2002) on the basis of an in-depth analysis of 17 left-turn collisions: they showed that in left-turn collisions, some of the problems are related to unconscious, attention errors when approaching and entering an intersection. Environmental factors such as uneven views when one approaches an intersection appear to exacerbate the problem. Accordingly, we could conclude that better signalized intersections could be a short-term solution for these types of accidents. In the long term, after taking into account additional information such as traffic circulation, one could consider the use of roundabouts to increase traffic safety in these black zones.

A second characteristic for accidents occurring in black zones are the rainy conditions. We know from other studies that there is a marked but complex relationship between the incidence of weather hazards and road accidents reported under such conditions (Edwards, 1996). Event if obvious, there is however no simple cause and effect relationship. It has already widely been demonstrated that during wet conditions accident numbers increase (see e.g. Brodsky and Hakkert, 1988). This is partly due to the slippery roads, but also to the fact that recent vehicle technical improvements (anti-lock brakes, four-wheel drive, traction control) have improved vehicle handling in poor conditions. Hence, drivers may take greater risks than they might have done otherwise, as they feel more confident when driving vehicles equipped with these safety features (Edwards, 1996). This phenomenon is also known as 'risk homeostasis' or 'moral hazard'. Table (4.13) mentions rain as well as aqua planing and a wet road surface [7,8]. This result is not surprising since aquaplaning can only occur when the road surface is wet! We kept all 3 attributes into account as 'rain' refers to weather conditions and 'wet road surface' to road surface conditions: an accident may occur after a shower, under fine weather conditions but with a still wet road surface. These aquaplaning patterns do not occur frequently when accidents are reported outside black zones and one should investigate whether and why the infrastructure characteristics do not prevent aquaplaning in black zones. In black zones, wet conditions seem to be not manageable by road users, for one reason or another (infrastructure related and/or behavioral).

Next, table (4.13) shows that accidents involving a pedestrian inside a built up area are also typical for black zones [9,10]. These collisions inside the built up environment frequently involve young road users (0_17 years) [11]. This probably indicates that black zones often correspond to road segments close to schools, playgrounds or other activities characterizing densely built areas. This confirms former papers showing that pedestrian injury collisions often occur when and where large numbers of pedestrians travel within complex roadway systems with high traffic flows (see e.g. Baker, Waller and Langlois, 1991; Braddock et al., 1994; LaScala et al., 2000 or Julien and Carr, 2002). Educational as well as environmental prevention efforts should hence focus on the harmonization of the road function and aspects such as traffic flow and local neighborhood as well as raising community awareness about the risks associated with them. This is especially true in Brabant Wallon (peri-urban) and for numbered roads: this road type is a relevant risk factor for pedestrians especially outside large urban centers where separate paths, sidewalks and pedestrian crossing with traffic lights are rare.

Furthermore, for some hazardous road segments, road users frequently lose control over the steering wheel to the left and hit a crash barrier as a result [12]. As expected, this type of accident often involves just one road user [12,13], and it also frequently takes place on a road with a speed limit of 120 km/hour [14,15] and less frequently on a road with a speed limit of 50 kilometers an hour [16]. Given the 120 km/hr limit, these accidents are located on highways. Run-off-roadway accidents have already often been studied (see e.g. literature review in Lee and Mannering, 2002). They are often related to an inadequacy of the speed and/or behavior of the user to the driving circumstances. The problem is then to identify cost-effective countermeasures that improve highway designs by reducing the probability of vehicles leaving the roadway and the severity of accidents when they do (roadside features). Lee and Mannering (2002) show that run-off-roadside features is a complex interaction of roadside features such as the presence of guardrails, miscellaneous fixed objects, sign supports, tree groups and utility poles along the roadway.

In-depth analysis of each sub-type of accidents should increase the understanding of each type of circumstances. This is, however, beyond the scope of this exploratory research. However, all these associations show that black zones often correspond to places where improvements could be made in terms of road design, signalization and better land-use planning. This corroborates other studies about road accidents and road geometry (see e.g. Agent and Deen, 1975; Wong and Nicholson,1992; Taber, 1998; Martin, 2002; Greibe, 2003).

Note that the accident patterns described in this section are limited to patterns that are discriminating black zones from non-black zones. More specifically, these accident patterns occur frequently inside black zones while they do not occur at all outside black zones. Hence, not all accident patterns that are characteristic for black zones are put forward in this section since they will not be able to uniquely describe the accidents in black zones.

4.7.4.2 Accident Patterns outside Black Zones

Although we are mainly interested in profiling black zones, describing the frequent accident patterns outside black zones can also give some useful information on the understanding of the spatial occurrence of traffic accidents. Therefore, we select the frequent item sets that are unique for accidents occurring outside a black zone and with very strong lift values. This results in 8 item sets of size 2 (lift <0,5 or lift >5), 84 item sets of size 3 (lift <0,5 or lift >5) and 238 item sets of size 4 (lift <0,5 or lift >15). Table (4.14) gives an overview of the most interesting of these frequent item sets.

In contrast to the results for accidents in black zones, table (4.14) shows that when an accident occurs outside a black zone, it frequently occurs on an intersection covered by traffic signs while making a left turn [17-19]. More specifically, these intersections are frequently located on roads with a speed limit of 50 kilometers an hour where no priority is given, resulting in a side impact collision [20-22]. In combination with the results of table (4.13), these patterns indicate that there exists a strong difference

Table 4.14: Frequent item sets for accidents outside black zones

N	Item 1	Item 2	Item 3	Item 4	sup	L
17	intersection traffic signs	left turn			7.4%	2.09
18	near intersection	intersection traffic signs	left turn		7.4%	12.42
19	2 road users	near intersection	intersection traffic signs	left turn	6.5%	20.48
20	speed 50 KMH	no priority	intersection traffic signs		5.6%	6.45
21	speed 50 KMH	near intersection	no priority	intersection traffic signs	5.6%	27.59
22	speed 50 KMH	sideways collision	near intersection	intersection traffic signs	6.1%	20.41
23	head on collision	negative way			8.6%	1.65
24	positive way	negative way	wet road surface		6.1%	0.49
25	car	positive alcohol test	drunken road user		6.9%	7.42

in the type of intersection on which the accidents take place, depending on whether the accident occurred inside or outside a black zone. This difference in type of intersections could also be explained by the traffic intensity on these locations. When there is less traffic, the probability of an accident is smaller and these locations are accordingly less often included in a black zone. Less traffic also means less public expenditures and hence less often modern traffic lights or roundabouts.

Next, table (4.14) also shows that the head-on collision is a frequently occurring accident type outside black zones with one road user driving in a negative way (related to the hectometer mark) [23]. Again, this pattern is not very surprising, since head-on collisions coincide with 2 road users driving in an opposite direction. This type of collision occurs frequently outside black zones while it does not appear at all as a frequent item set in the results of the previous section. Furthermore, no other association concerning this type of collision was clearly put forward. Hence, it indicates that risk-taking may play a dominant role in head-on collisions. This confirms former results (see e.g. Rajalin, 1994; Larsen and Kline, 2002). In-depth analysis of the accidents is here needed to further understand the frontal accidents. Additionally, with one road user driving in a positive way and another road user driving in a negative way, the accident will less frequently than expected take place on a wet road surface [24]. These results can probably be explained by the fact that aquaplaning does not frequently occur outside black zones (see previous section). Accordingly, head-on accidents are less frequently related to rainy weather but more to risk-taking behavior on the road leading to spatially scattered frontal accidents.

Finally, table (4.14) also indicates that accidents where at least one car is involved outside black zones frequently involve a drunken road user and a positive alcohol test [25]. This pattern is not very surprising, since a positive alcohol test is an indication for a drunken road user. However, we should bear in mind that this accident pattern does not emerge for black zones. A possible explanation could be that outside black zones accidents are less related to infrastructure characteristics but more to behavioral aspects such as drinking and driving. Accidents are then more scattered and hence spatially occur at random. There is indeed a priori no reason for alcohol related accident to be more clustered than others. Again, further research is needed to confirm or invalidate the results put forward in this analysis.

4.7.4.3 Common Accident Patterns

In the previous sections, we respectively discussed the accident patterns for accidents occurring inside black zones and outside black zones. In this section, a number of frequent item sets are discussed that occur in both data sets and accordingly they describe hazardous as well as non hazardous road segments. However, the occurrence of these patterns will not be equally as strong in both data sets. Therefore, we can use the interest value Int (see definition (4.2)) to identify the accident patterns that occur more frequently inside than outside the black zones.

Accordingly, selecting the item sets with Int > 0.5 resulted in 14 item sets of size 2, 208 item sets of size 3 and 167 item sets of size 4. However, at the same time, we will use the lift values to determine the item sets that are not only discriminating between black zones and non black zones but that are also very descriptive. Selecting these item sets resulted in 14 item sets of size 2, 9 item sets of size 3 (lift>2) and 19 item sets of size 4 (lift>2).

Table (4.15) gives an overview of the most interesting frequent item sets. Note that the lift values in this table correspond to the values for the black zones, since these are

Table 4.15: Frequent item sets inside and outside black zones

N	Item 1	Item 2	Item 3	Item 4	s_b	L_b	s_n	L_n	Int
26	night	highway			12.3%	1.39	5.6%	1.39	0.54
27	public lighting	night	road with separated lanes		13.4%	3.78	6.2%	3.77	0.54
28	highway	weekend			16.8%	1.15	7.4%	1.15	0.56
29	parallel collision	highway			13.4%	1.18	6.3%	1.18	0.53
30	parallel collision	road with separated lanes	highway		13.4%	2.09	6.1%	2.09	0.54
31	rain	120 KMH			10.8%	1.09	5.0%	1.09	0.53
32	wet road surface	road with separated lanes	rain		17.9%	2.66	5.9%	2.66	0.67
33	wet road surface	outside built up area	road with separated lanes	rain	16.9%	3.73	5.7%	3.74	0.66
34	car	wet road surface	road with separated lanes	rain	5.1%	3.04	17.2%	3.04	0.70
34	car	public lighting	road with separated lanes	rain	14.2%	2.34	5.2%	2.34	0.63

the main interest of this research. This table shows that accident patterns that are typical for all accidents, and that also have a high interestingness value (pointing out that they occur more frequently inside than outside black zones) are mostly related to accidents on highways or roads with separated lanes. More specifically, these accidents occur more frequently than expected at night [26,27]. This confirms the results of other researchers that showed that accident rates vary on different types of roads depending on day and night conditions (see e.g. Martin, 2002). Furthermore, these accidents frequently take place during weekends [28]. Once again, previous studies have demonstrated that the time of the day and week is an important risk factor, especially for young drivers (see e.g. Doherty et al., 1998). The type of accident is often a parallel collision [29,30]. These accident patterns respectively occur in more than 12% (Sb=12.3%) [26], 13% [27], 16%[28], 13% [29] and 13% [30] of all accidents in black zones while outside black zones these types of accidents only occur in approximately 5% (Sn=5,6%) [26], 6% [27,29,30] or 7% [28] of all accidents.

Furthermore, accidents in black zones are more often related to the rainy weather and/or a wet road surface than accidents that take place outside black zones. More specifically, these accidents frequently occur on roads with separated road lanes [31-35] with a speed limit of 120 kilometers an hour [30] and outside the built up area [31]. These results can be related to the figures in table (4.13), which indicate that aquaplaning is an important problem in black zones. Since we can assume that (for this empirical study) the weather conditions and accordingly the amount of rain inside black zones and outside black zones will not be significantly different, accidents caused by rain will obviously also play an important role outside black zones, explaining the occurrence of this accident factor in both data sets. However, as explained before, wet conditions are less manageable by road users in black zones, for one reason or another (infrastructure related or behavioral).

4.7.5 Conclusions

In this research, we aimed at understanding why road accidents tend to cluster in specific road segments. More particularly, the technique of frequent item sets is applied for automatically identifying accident circumstances that frequently occur together, for accidents located in and outside black zones. Our analysis here is limited to one administrative region: the Walloon Brabant, for the period 1997-1999.

The most important results of this research are that road accidents concentrations in black zones correspond to specific frequent items. Taking a left turn is an important accident factor as well inside as outside black zones. However, in black zones, these accidents frequently take place on intersections covered by traffic lights while outside black zones, this accident type frequently occurs on intersections with traffic signs, which could be explained by the traffic on these accident locations. Better signalized intersections could be a short-term solution for these types of accidents, however, in the long term, one should also consider the use of roundabouts to increase traffic safety in these black zones.

A second important accident circumstance as well inside as outside a black zone is the rainy weather conditions. However, inside black zones this factor frequently coincides with aquaplaning, which is not the case outside these black zones. These results suggest that black zones and non black zones are characterized by different infrastructure specifications, explaining the occurrence of the clustering of accidents in black zones.

Furthermore, a collision with a pedestrian involving young road users inside the built up area is a typical accident pattern that frequently occurs inside a black zone. This confirms common sense and former papers showing that pedestrian injury collisions often occur when and where large numbers of pedestrians travel within complex roadway systems with high traffic flows. Education and environmental prevention efforts should hence focus on aspects of traffic flow as well as raising local neighborhood community awareness about the risks associated with them.

Additionally, loss of control over the steering wheel and the resulting collision with a crash barrier is a frequently occurring accident pattern in black zones. These run-off-roadway accidents often occur on freeways and are related to an inadequacy of the speed and/or behavior of the user to the driving circumstances. The problem is then to identify cost-effective countermeasures that improve highway designs by reducing the probability of vehicles leaving the roadway and the severity of accidents when the do (roadside features).

The findings of this research are rather suggestive and limited to one data set. Furthermore, in this research the location and the length of the black zones are defined by means of local spatial autocorrelation indices and are considered as given in our problem. Results of this research show, however, the usefulness of the frequent item sets in analyzing the combination of patterns associated with road accidents occurrences in these black zones. More specifically, these results show that a special traffic policy towards accidents in black zones and accidents outside these zones should be considered. Indeed, these spatial concentrations of accidents are characterized by specific accident circumstances, which require different countermeasures to reduce their number such as improvements in terms of road design, signalization, and local environment. Accordingly, infrastructure and land-use can enhance traffic safety but is not an answer to all problems. Finally, one should also mention that there is no unique combination of characteristics associated to road accident occurrences: it is a complex phenomenon for which only some aspects are reported here.

4.8 Profiling High Risk Road Clusters

In this empirical study, we will illustrate the possibility of identifying geographical locations with high accident risk by means of clustering techniques and profiling them in terms of accident related data by means of data mining techniques using a small but complex data set of traffic accidents.

In particular, in the first part of this research, we will use model based clustering to cluster traffic roads into distinct groups based on their similar accident frequencies. In the second part of this research, the data mining technique of frequent item sets is used to profile each cluster of traffic roads in terms of the available traffic accident data.

Analogously with the previous empirical studies, the data for this research originate from the National Institute for Statistics and are obtained from the Belgian 'Analysis Form for Traffic Accidents' that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. More specifically, this analysis will focus on 19 central roads in the city of Hasselt for three consecutive time periods of 3 years each: 1992-1994, 1995-1997, 1998-2000. In total, 142 accidents are included in the analysis.

4.8.1 Clustering Traffic Roads

4.8.1.1 Model Formulation

As explained in the previous section, the number of accidents on 19 ($n = 19$) similar roads in Hasselt (Belgium) are considered for three following time periods of each three years. The idea is to cluster the traffic roads in groups based on the similarities in the number of accidents that occurred on the roads during each time period.

Therefore, a 3-variate Poisson distribution (Y_1, Y_2, Y_3) with one common covariance term is defined for each cluster (Li et al., 1999):

$$\begin{aligned} Y_1 &= X_1 + X_{123} \text{ (period1} = 1992 - 1994) & (4.10)\\ Y_2 &= X_2 + X_{123} \text{ (period2} = 1995 - 1997) & (4.11)\\ Y_3 &= X_3 + X_{123} \text{ (period3} = 1998 - 2000) & (4.12) \end{aligned}$$

with Y_i = the number of accidents on a traffic road in period i and all X's independent univariate Poisson distributions with respective parameters ($\lambda_1, \lambda_2, \lambda_3$, λ_{123}).

Since a large number of variables that influence the number of road accidents in a certain time period will be time specific (e.g. traffic intensities), we will use one Poisson distribution for each time period to approximate the number of accidents in period i (X_i). Furthermore, it can easily be seen that the occurrence of accidents on a traffic road over several time periods may be related (e.g. due to bad infrastructure). Therefore, correlations between the observations in each cluster are allowed by identifying the parameter $\lambda_1 23$, which can be considered as a covariance factor that measures the risk of the area common to all time periods (Karlis, 2000).

4.8.1.2 Model estimation

The algorithm is sequentially applied to the data for 1 to 5 clusters (k =1,...,5). Furthermore, in order to overcome the dependence on the initial starting values for the model parameters, resulting in a local optimum instead of a global optimum value, different sets of starting values for p_i and λ_i are chosen. However, results show that dependencies on the initial starting values only occur for large values of k, while for smaller values of k the algorithm terminates at the same solution with the same parameter values, indicating that the global optimum has very likely been achieved.

Figure (4.6) shows the evolution of respectively the loglikelihood and the information criteria for different clusters (k=1,...,5) of the 3-variate Poisson Mixture Model with common Covariance. Note that the depicted values for the AIC, BIC and CAIC are re-scaled in order to be comparable to the loglikelihood. This figure indicates the use of the goodness of fit measures to determine the number of clusters: although the loglikelihood of the model increases when the number of clusters increases, the information criteria will not choose the maximum possible clusters to cluster the data. Considering the model complexity, the AIC selects three clusters whereas the CAIC and the BIC select only two clusters. This difference can be explained by the fact that the AIC does not consider the size of the data set, whereas the CAIC and the

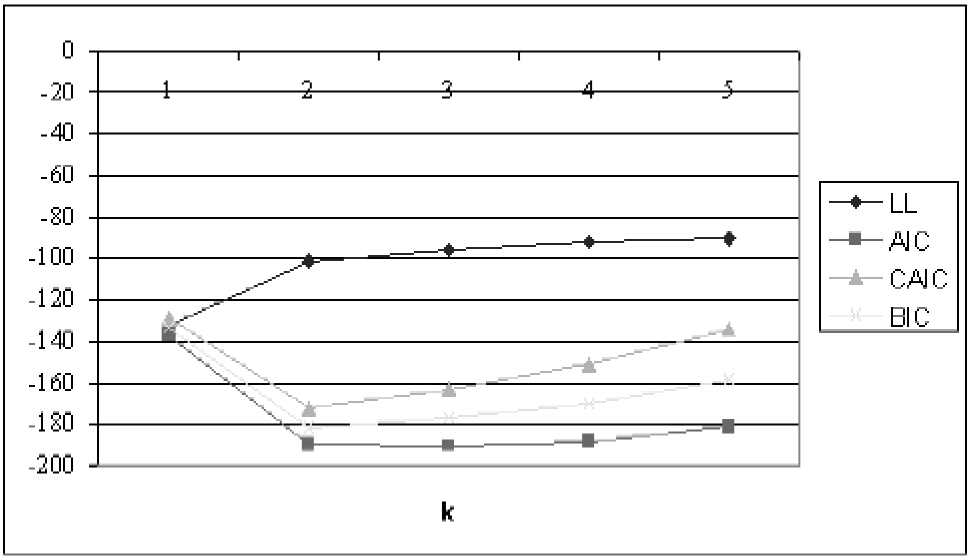

Figure 4.6: LL, AIC, BIC and CAIC against the number of clusters k

BIC do penalize for this factor. However, note that the difference between the AIC value for two clusters and for three clusters is very small.

For the remainder of this research, we will focus on the results of the 2-components common covariance model, which groups the traffic roads in two clusters.

4.8.1.3 Parameter Estimates

Table (4.16) contains the parameter estimates and the size of each cluster (p) for the model with 2 clusters.

In this model, the average number of accidents increases per period for the first cluster and decreases per period for the second cluster. Furthermore, the observed average accident rate per period for cluster 1 is mainly dependent on the average

Table 4.16: Parameters for the 2-components model

	Parameters				
Cluster	λ_1	λ_2	λ_3	$\lambda_1 23$	**p**
1	0.631	0.930	1.089	0.005	0.688
2	4.149	3.490	1.790	3.726	0.312

accident frequency of the concerning period (λ_i) and less on the covariance factor (λ_{123}). For cluster 2, the covariance term does play an important role in the observed average accident rate per period. This can be explained as for this cluster there is a strong common factor in all periods that has to do with the accident risk on these roads, for example due to bad infrastructure, constant high traffic volume. However, note that for this second cluster there is a strong decrease in the average number of accidents from period 2 (λ_2) to period 3 (λ_3). This could be an indication of infrastructure changes between these two periods.

4.8.2 Profiling Traffic Roads

4.8.2.1 Preprocessing the Data Set

In the second part of this research, we will use frequent item sets to profile the two clusters of traffic roads. Therefore, two data sets of traffic accidents are defined according to the traffic roads belonging to cluster 1 and cluster 2. This is determined by estimating the posterior probability w_{ij}, i.e. the posterior probability for location i to belong to cluster j. This probability can be obtained for each observation vector y_i according to Bayes' rule (see equation 3.5). Indeed, after estimation, we know the density distribution f($y_i|\theta_j$) with θ_j is vector of parameters for cluster j, and we know the cluster size p_j of each component such that we can calculate the posterior distribution as follows:

$$w_{ij} = \frac{p_j f(y_i|\theta_j)}{\sum_{j=1}^{k} p_j f(y_i|\theta_j)} \tag{4.13}$$

Assigning the accident locations to the cluster with the highest posterior probability resulted in a total of 35 traffic accident records that were included for the analysis of cluster 1 (13 traffic roads) and 107 traffic accidents that were included for the analysis of cluster 2 (6 traffic roads). Figure (4.7) gives an overview of these clusters.

After preprocessing the available data to prepare them for the use of the mining technique (discretize variables and remove irregularities such as data inconsistencies, missing values, redundant variables and double counts), two steps can be distinguished in the mining process: a mining step for generating the frequent item sets and a post-processing step for evaluating and interpreting the most interesting patterns.

4.8.2.2 Generating Frequent Item Sets

A minimum support value of 30% was chosen for the analysis by means of frequent sets. From cluster 1, the algorithm obtained 29,415 frequent item sets of maximum size

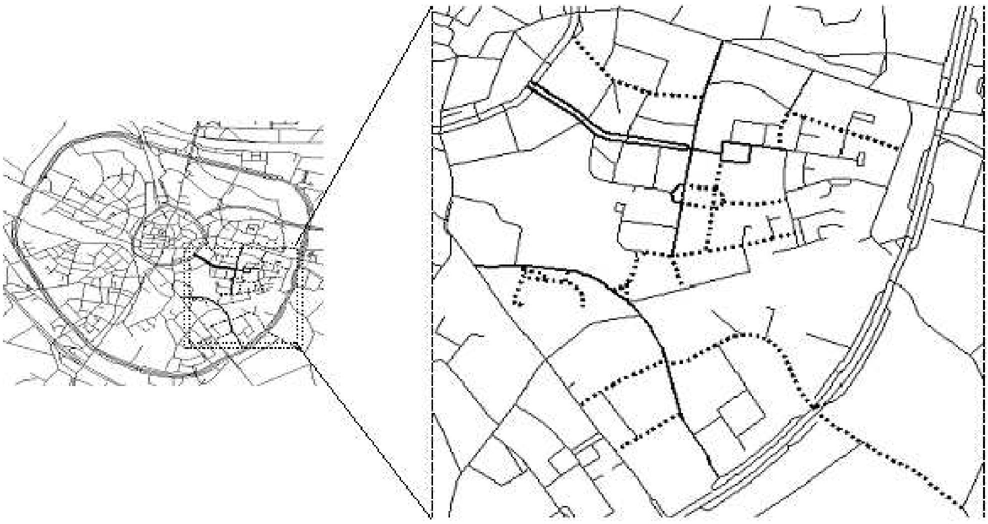

Figure 4.7: Traffic roads belonging to cluster 1 (...) and cluster 2 (in bold)

4. Although these results relate to a relatively small number of accident records, they are quite reasonable since an average of 40 items is available per accident, allowing the algorithm to generate multiple combinations of size 4 item sets. With the same support parameter the second analysis resulted for cluster 2 in 28,541 frequent item sets of maximum size 4. These rules are further processed to select the most interesting rules.

4.8.3 Results

4.8.3.1 Common Accident Patterns

First, we will discuss the frequent item sets that are descriptive for both clusters of accident locations. Selecting these frequent item sets resulted in 24,562 accident patterns. However, in this research, we will pay special attention to the item sets with a positive interest value, i.e. approximating '1' since these accident patterns are stronger for cluster 2, i.e. the cluster with the highest accident risk. Accordingly, selecting the item sets with Int > 0.3 resulted in 12 item sets of size 2, 75 item sets of size 3 and 309 item sets of size 4. Table (4.17) gives an overview of the most interesting of these frequent item sets.

These results show that the accident patterns that occur more frequently in cluster 2 than cluster 1 often occur on a weekday, inside the built up area with 2 road users

[N=1], with one road user's age being between 18 and 29 [2]. These accidents often involve a female road user [3], driving a car [4].

Furthermore, one road user is frequently driving in a straight direction with a speed limit of 50 kilometers per hour in daylight [5]. These accidents frequently result in a sideways collision [6,7,8] with at least one road user in a normal condition [9]. The relatively young age (between 18 and 29) of the road users involved in the accidents on the traffic roads of cluster 2 could indicate that in this high risk cluster young parents are involved who drive their children to school. Indeed, schools are one stable geographic feature associated with regular, often concentrated periods of complex and congested traffic patterns (LaScala et al., 2004). However, additional information on the traffic roads shows us that the schools are located on the traffic roads of cluster 1. On the traffic roads of cluster 2 the church, music academy, cultural center and most important shops (e.g. bakeries, newspaper shops, hairdressers, laundry services, banks) are located. This information indicates that these accident patterns do not occur in the immediate environment of the schools but will most likely occur on the roads leading to and from the schools where the shops are located. Indeed, when looking at the time of accident, 14.5% of the accidents take place between 7am and

Table 4.17: Frequent item sets for both clusters

N	Item 1	Item 2	Item 3	Item4	s_2	L_2	s_1	L_2	Int
1	weekday	Inside built up area	2 road users		0.72	1.02	0.45	0.94	0.37
2	weekday	age 18-29			0.59	1.02	0.39	0.92	0.35
3	weekday	inside built up area	female		0.52	1.01	0.32	0.99	0.38
4	weekday	inside built up area	female	car	0.54	1.01	0.32	0.75	0.40
5	straight direction	50 KMH	daylight		0.47	1.06	0.32	1.10	0.30
6	sideways collision	50 KMH			0.49	1.01	0.32	1.05	0.33
7	sideways collision	inside built up area	weekday	car	0.58	1.09	0.39	1.03	0.33
8	sideways collision	inside built up area	weekday	2 road users	0.58	1.09	0.39	1.03	0.33
9	sideways collision	female	normal condition		0.55	1.02	0.35	0.88	0.35
10	dry road surface	inside built up area	weekday		0.58	0.99	0.39	0.84	0.33
11	normal weather	inside built up area	weekday	straight direction	0.52	0.97	0.32	0.74	0.38

9am and 18.3% of the accidents take place between 4pm and 6pm. Furthermore, 29% of the accidents take place between 10am and 12am and 23.7% between 1pm and 3pm, which correspond with the opening hours of most shops. Remarkably fewer accidents occur in the evening (5,3% between 7pm and 9pm) and at night (9.2% between 10pm and 6am) when the shops are closed. Additionally, the accidents occurring on a weekday, inside the built up area and resulting in a sideways collision mostly occurred on a crossroad (71.6%).

Additionally, results of table (4.17) show that compared to cluster 1 the accidents on traffic roads belonging to cluster 2 occur more frequently on a dry road surface [10] and with normal weather [11]. However, note that for both clusters these accident patterns have a lift value smaller than '1'. This means that although in cluster 2 more accidents occur under normal weather conditions and on dry road surfaces than in cluster 1, these accident patterns still occur less frequently than expected for both clusters. All these accident patterns indicate that most accidents that occur on a traffic road belonging to the high-risk cluster (cluster2) take place under no special variable environmental circumstances (e.g. rain, alcohol).

Therefore, it can be expected that the high number of accidents on these traffic roads can be explained by an unsafe infrastructure or a high traffic volume for all time periods, confirming our previous results of a high common covariance factor for this cluster. Indeed, several authors have described the relationship between accident involvement on the one hand and road geometry and traffic volume on the other hand (see e.g. Abdel-Aty and Radwan, 2000).

4.8.3.2 Accident Patterns for High Risk Cluster

Next, we will discuss the accident patterns that are frequent for cluster 2 but not for cluster 1. These item sets represent very characteristic combinations of accident circumstances for the traffic roads with a high accident risk. More specifically, we are interested in the frequent item sets with lift values differing from '1' since these item sets represent strong dependencies between the different items of the item set. However, note that we should not compare the absolute lift values of the item sets of different sizes, since the more items the item set consists of, the higher the lift value will become.

Selecting the item sets that are unique for cluster 2 resulted in 3,943 frequent accident patterns (table (4.18)). Conform with the previous results, the results of table (4.18) show that sideways collisions involving female road users are a typical

Table 4.18: Frequent item sets for the high risk cluster

N	Item 1	Item 2	Item 3	Item4	sup	L
12	sideways collision	female	no priority		0.41	1.23
13	sideways collision	female	no priority	crossroad	0.35	1.38
14	50 KMH	brakes	no priority		0.31	1.40
15	50 KMH	car	age 18-29		0.39	1.19
16	weekday	bicycle			0.34	1.09
17	weekday	bicycle	2 road users		0.31	1.14
18	0 deadly injured	bicycle			0.36	1.13

accident pattern for traffic roads with a high accident risk [12,13]. Again, these results indicate that this type of accident occurs frequently while the maximum speed limit was 50 kilometers per hour for these accidents, while no priority is given [12,13,14] and the age of at least one road user was between 18 and 29 [15].

A second important accident type that is reflected in the results of table (4.18) are the accidents involving a bicycle. These accidents often take place on a weekday [16] with 2 road users [17] and frequently coincide with 0 deadly injured victims [18]. Note that these accident patterns are not very surprising as such, but remark that they do not appear for the accidents of cluster 1. Again, this could be explained by the proximity of shops, the cultural center, music academy, church etc. on the traffic roads of cluster 2. The intensity of bicyclists will probably be much higher on these roads compared to majority of the roads of cluster 1 where in general none of these centers or stores is located. Since this traffic intensity of bicyclists will more or less be the same over all time periods, this factor will probably contribute to the high common covariance term for cluster 2. However, since the schools are located on the traffic roads of cluster 1 and accordingly the intensity of bicyclists will also be high in these specific streets, it is surprising that these accident patterns with bicyclists do not occur at all in cluster 1. Again, this indicates that accidents with bicycles do not frequently occur in the immediate environment of the schools but will most likely occur on the roads leading to and from the schools.

4.8.3.3 Accident Patterns for Low Risk Cluster

Finally, we will discuss the item sets that are unique for the accidents related to cluster 1. These item sets represent very characteristic combinations of accident circumstances for the traffic roads with a low accident risk. Again, we are interested

Table 4.19: Frequent item sets for the low risk cluster

N	Item 1	Item 2	Item 3	Item4	sup	L
19	crossroad	priority to the right			0.62	1.40
20	crossroad	priority to the right	daylight		0.42	1.47
21	normal weather	priority to the right	age 30-45		0.32	1.16
22	normal weather	dry road surface	crossroad	age 46-60	0.35	1.07
23	car	age 46-60			0.35	1.07
24	inside built up area	weekend			0.32	1.14

in the frequent item sets with lift values differing from '1'. Selecting the item sets that are unique for cluster 1 resulted in 4,879 frequent accident patterns. The most interesting of these frequent item sets can be found in table (4.19).

These results show that an important accident type for the traffic roads with low accident risk are the accidents on crossroads with priority to the right [19,20]. These accidents take up 61,61% of all accidents on these roads. However, in contrast with the previous results, these accidents more frequently than expected involve a road user with age between 30 and 45 [21].

Additionally, when an accident occurs on a crossroad with normal weather on a dry road surface, at least one road user of the age between 46 and 60 is involved [22,23]. These results show that the age of the road user is not as pronounced for the accidents occurring on the low accident risk traffic roads.

Finally, an important accident pattern involves the accidents that occur inside the built up area in the weekend [24]. Note that these weekend accidents did not appear for the traffic roads with high accident risk.

4.8.4 Conclusions

In the first part of this research, model based clustering is used to cluster traffic roads into two groups based on their similar accident frequencies. Results show that the observed average accident rate per period for cluster 1 is mainly dependent on the average accident frequency of the concerning period and less on the covariance factor. For cluster 2, the covariance term does play an important role in the observed average accident rate per period. This can be explained as for this cluster there is a strong common factor in all periods that has to do with the accident risk on these roads.

In the second part of this research, the association algorithm was used on a data set of traffic accidents to profile the two clusters of traffic roads. Results showed that, in contrast with the results for cluster 1, the accident patterns for cluster 2 confirm that the roads belonging to this cluster should be considered as dangerous at all times resulting in a high number of accidents in all time periods. Furthermore, a special traffic policy towards these clusters should be considered, since each cluster is characterized by specific accident circumstances.

4.9 Discussion and limitations

In this chapter, the association algorithm was used on a data set of road accidents to profile hazardous accident locations in terms of accident related data and location characteristics. More specifically, frequent item sets and association rules are generated to identify accident circumstances that frequently occur together. The use of this technique is of an explorative character since it describes the co-occurrence of accident circumstances but it does not give any explanation about the causality of these accident patterns. Therefore, its role is to give direction to more profound research on the causes of these accident patterns and explanation which will require the use of some additional techniques or expert knowledge. These patterns represent interesting interactions in accident factors, which accordingly can be used to test in statistical models. Furthermore, the use of frequent item sets not only allows to give a descriptive analysis of accident patterns and discern different accident types, it also creates the possibility to find the accident characteristics that are discriminating between groups such as high frequency accident locations and low frequency accident locations, black zones and non black zones or clusters of traffic roads.

However, although the analyses carried out in this research revealed several interesting patterns which, in turn, provide valuable input for purposive government traffic safety actions, some final remarks are in place. First, according to Elvik (2006), the role of accident analysis should also be to identify false positives, these being black spots at which no clearly interpretable pattern in accidents can be found. Indeed, no way of identifying black spots is perfect, so there will always be a few false positives amongst the black spots identified. Accordingly, the author proposes a new approach to the analysis of accidents at hazardous road locations, designed to make these analyses more effective in discriminating between true and false positives, i.e. sites whose true expected number of accidents is high and sites where a high recorded number of accidents is predominantly the result of chance variation. However, since in this research, we use data mining to find accident patterns that frequently occur at an aggregated level of accident locations and not at the level of the individual location, it is impossible to single out these false postives. Therefore, no conclusions were made on the existence of false positives among the identified hazardous locations and this techniqe should merely be seen as an exploratory tool to identfy accident patterns for a specific group of accident locations such as black spots.

Next, the skewed character of the accident data limits the amount of information contained in the data set and will therefore restrict the number of circumstances that will appear in the results. Moreover, the choice for the minimum support parameter can prevent the association algorithm from generating rules on the less frequent accident conditions. However, this information on rare accident types could be very useful since the circumstances in which these accidents occur, will probably not be trivial and more difficult to discern.

Additionally, the inclusion of domain knowledge (e.g. traffic intensities, a priori infrastructure distributions) in the association algorithm would improve the mining capability of this data mining technique and would facilitate the post-processing of the association rules set to discover the most interesting accident patterns. Moreover, the variables used here are restricted to those collected by the police. Ideally, our range of variables should therefore be extended to other sources (traffic, land-use, accessibility, etc.) in order to better approach the explanation process.

Finally, considering the large number of attributes in the traffic accident data set, it seems interesting to explore the potential of techniques that generate rules with longer patterns to uncover more complex associations in traffic accidents. Indeed, in this research, using the apriori algorithm and the chosen minimum support values, we were only able to generate accident patterns of maximum size 4. Accordingly, in order to generate all frequent item sets in the data, we adopted an algorithm of Grahne and Zhu (2003), who focussed their work on the mining of frequent item sets in large databases. More specifically, the authors introduced a novel technique to mine for frequent item sets by using an array to greatly improve the performance of the algorithms operating on FP-trees. As a test case, we used this algorithm to generate all frequent item sets for the accidents that occurred inside and outside black zones (see section (4.7)). Table 4.20 gives an overview of the results when using the same minimum support value of 5%.

Table 4.20: All item sets for accidents inside and outside black zones

	Number of Item Sets	
Size of the Item Set	**Inside Black Zones**	**Outside Black Zones**
<=4	181,761	181,066
5	585,012	527,506
6	1,503,629	1,303,513
7	2,821,858	2,327,294
8	3,985,326	3,061,149
9	4,299,310	2,955,907
10	3,555,297	2,060,841
11	2,246,180	1,015,737
12	1,077,804	346,898
13	390,096	80,945
14	105,006	12,776
15	20,221	1,342
16	2,555	85
17	177	2
18	4	0

These results show that, indeed, there exist long patterns in the accident data (maximum size of the item set = 18). On the one hand, these patterns could reveal some new and surprising information on the accident data given the large number of accident characteristics that are combined. On the other hand, the size of these item sets also implies that the patterns will be more difficult to understand. This will, in turn, render it more difficult for policy makers to find an explanation and decide on suitable actions to prevent these accident patterns. Additionally, note that although the accidents that occurred inside black zones represent only 30.05% of the total number of accidents (553/1840), the number of frequent item sets that are generated for this group of accidents is almost equally as large as for the accidents outside black zones. In other words, although this group of accidents is much smaller, an equal number of accident patterns is found to describe these accidents, indicating the complexness of accidents occurring inside black zones.

Given this information and the fact that we are mainly interested in accident patterns inside black zones, we examined into more detail the item sets of size 8 for the black zone accidents. We compared these item sets with the results of section (4.7.4.1) to find out whether the increased complexity of these longer patterns is compensated by the more explanatory power of these item sets.

First, we selected the item sets of size 8 with the highest lift values. More specifically, we looked at the 110 item sets with a lift value > 80 and a support value of approximately 5% (remember that the more items the item set consists of, the higher the lift value will become). However, results showed that these item sets describe the same accident patterns as the 4-item sets that we discussed in section (4.7.4.1). The additional items in the item sets do not give more surprising information on the circumstances in which these accident patterns occurred. Next, given the relatively low support value of the 110 item sets discussed above, we also selected the item sets of size 8 with the highest support values. More specifically, we selected the item sets with a support value of 25% and a lift value >=2. However, as the the relatively small lift values indicate, these item sets do occur frequently but do not reveal any surprising or interesting accident patterns.

In conclusion, we can state that although it is possible to generate a large number of item sets with more than 4 items, for this research the increased complexity of these longer patterns is not compensated by the more explanatory power of these item sets.

Chapter 5

Evolution of Dangerous Accident Sites over Time

In this section, dangerous accident locations based on the data from 1997-1999 are compared with the accidents that should be considered as dangerous based on the data from 1999-2001. More specifically, we will we will apply the currently used ranking and selection criterion in Flanders on the data from 1999-2001 to investigate whether dangerous accident locations, based on the priority values, tend to migrate over time in the region of Flanders.

5.1 Introduction

As explained in chapter 3 of this dissertation, the 1,014 accident locations that are currently considered as dangerous by the Flemish government are selected using the large data set of traffic accidents from the National Institute of Statistics for the region of Flanders (Belgium) for the period 1997-1999. To improve the traffic safety on these locations, the Flemish government will each year, starting in 2003 for a period of five years, invest 100 million EURO to redesign the infrastructure of the 800 accident locations with the highest score.

However, although the Flemish government is still focussed on tackling the dangerous accident sites of 1997-1999, the accident data of 1999-2001 are currently also available for analysis. Accordingly, instead of using the data from 1997-1999 and performing a sensitivity analysis on the currently used ranking and selecting criterion, in this chapter we use the data from 1999-2001 to make a comparison between the locations of the dangerous sites in 1997-1999 and 1999-2001 in Flanders. As explained in section 2.1.5, a relation has been observed between a hazardous accident site disappearance and the development of a new dangerous accident location near the previous one. This phenomenon of accident migration occurs when a dangerous accident site is not managed properly: on the one hand the accident number on the existing hazardous site decreases, while, on the other hand, usually in a road section nearby, the accident number suddenly increases. Accordingly, we will apply the currently used ranking and selection criterion in Flanders on the data from 1999-2001 to investigate whether the dangerous accident locations indeed tend to migrate over time in the region of Flanders by comparing the locations of the dangerous accident sites between the periods 1997-1999 and 1999-2001.

5.2 Comparative analysis

5.2.1 Mapping the Dangerous Accident Sites

In accordance with the 1,014 currently selected dangerous accident locations based on the data from 1997-1999, we used equation (3.1) to calculate the priority score for the accident sites of the accident data of 1999-2001. This resulted in 991 accident locations where at least 3 accidents occurred and with a priority score that equals 15 or more. Accordingly, these accident sites should be considered as dangerous for this period.

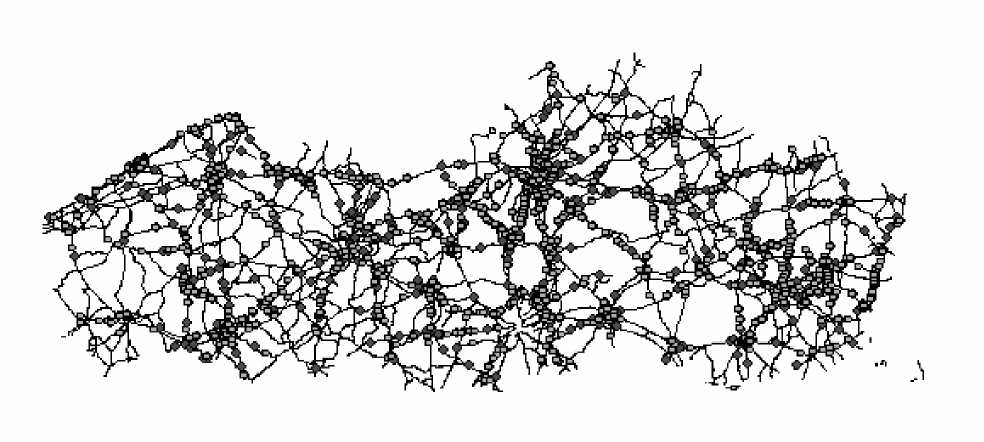

Figure 5.1: Dangerous accident locations Flanders 1997-1999

Figure 5.1 shows the 1,014 dangerous accident locations that are currently considered as dangerous based on the data of 1997-1999 (diamond shaped black dots) and the 991 accident sites that should be considered as dangerous based on the data of 1999-2001 (rectangular shaped grey dots).

Comparing these dangerous accident locations of 1997-1999 and 1999-2001 in a Geographical Information System (GIS) shows that the location of 315 dangerous accident sites is exactly the same in the two periods. This means that 31.07% of the 1,014 accident sites that are currently considered as dangerous should still be considered as dangerous based on the data from 1999-2001. Furthermore, this indicates that 68.21% of the dangerous accident sites from 1999-2001 were not considered as dangerous in 1997-1999.

However, although the identification of dangerous accident locations is in theory related to intersections with a radius of 50 meters or roadway segments of numbered roads with a length of 100 meters, this location information in GIS is collected on a more detailed level than the hectometer. Accordingly, due to improved measuring and calibration techniques the reference to the exact location of the dangerous sites can differ between the two periods. Consequently, by comparing these exact locations of the dangerous accident sites in the two periods, we obtain an underestimation of the actual number of dangerous accident sites that correspond between the two periods.

To correct for this underestimation, we rounded down the location data of the dangerous accident sites on road segments for the two periods to 100 meters. This corresponds with the definition of dangerous accident locations in Flanders. Comparing these location references resulted in 470 dangerous accident sites that correspond between 1997-1999 and 1999-2001. Analogously, 46.36% of the dangerous accident locations in 1997-1999 should still be considered as dangerous based on the data from 1999-2001. On the other hand, 47.43% of the dangerous accident sites from 1999-2001 were not considered as dangerous in 1997-1999.

5.2.2 Analyzing the Different Rankings

A closer investigation of the 470 accident sites that are common in the two periods shows that the ranking of these accident locations based on the data from 1997-1999 varies between 1 and 1,014. This means that the dangerous accident sites that are still dangerous based on the data from 1999-2001 are not only the 470 most dangerous ones.

More specifically, table (5.1) shows for different subsets of the 1,014 accident locations from 1997-1999 the number of accident sites that are also dangerous in 1999-2001.

When analyzing the top 15% most dangerous accident sites of 1997-1999 (152 sites), table (5.1) shows that 71.05% of these accident sites are still dangerous in 1999-2001. Even more, when closer investigating these sites, results show that 9 accident sites that were ranked in the top 10 of most dangerous accident locations in 1997-1999 are also dangerous in the period 1999-2001. For the top 40%, top 70% and top 800 these figures diminish respectively to 61.33%, 53.80% and 52.12% accident

Table 5.1: Dangerous accident sites in both periods

1997-1999	**1999-2001**
top 15% (152 sites)	71.05 %(108 sites)
top 40% (406 sites)	61.33% (249 sites)
top 70% (710 sites)	53.80% (382 sites)
top 800 (800 sites)	52.12% (417 sites)

Table 5.2: New dangerous accident sites

1999-2001	**1997-1999**
top 15% (149 sites)	24.16 % (36 sites)
top 40% (397 sites)	35.01% (139 sites)
top 70% (694 sites)	45.24% (314 sites)
top 800 (800 sites)	47.87% (383 sites)

sites that are dangerous in both periods.

Conversely, table (5.2) shows for different subsets of the 991 accident locations that should be considered as dangerous based on the data from 1999-2001, the number of accident sites that were not belonging to the most dangerous accident sites in 1997-1999.

Results from table (5.2) show that from the top 15% most dangerous accident sites in 1999-2001, 24.16% accident locations were not even considered as dangerous based on the data from 1997-1999. Additionally, the larger the subset, the higher the number of accident sites that are not considered as dangerous based on the data from 1997-1999, but should be considered as dangerous sites based on the data from 1999-2001 (respectively 35.01%, 45.24% and 47.87% for the top 40%, top 70% and the top 800).

5.2.3 Comparing Results with Bayes

In section 3.5.1, we explained that when identifying hazardous locations, the actual count of accidents are subject to random variation and to the regression to the mean problem. Accordingly, locations that in one period recorded 'x' accidents do not have, on the average, 'x' accidents in the subsequent period. By using Bayesian estimation values we could handle this uncertainty and the great variability of accident data and produce a probabilistic ranking of the accident locations.

Analogously with the research in section 3.5.1, we should take into account this variability in the accident data when comparing the rankings of the dangerous accident sites in 1997-1999 and 1999-2001. Indeed, as explained in section 2.1.2, it is possible that the high crash rates observed at some sites in 1997-1999 may be due to chance, or a combination of both chance and a moderately hazardous nature. In that case, a dangerous accident site may be nothing more than a meaningless cluster of accidents which occurs randomly and has no intrinsic meaning. These sites are likely to have fewer crashes in 1999-2001 because the number of crashes will tend to gravitate towards the long-term mean value.

Therefore, in this section, we use the probabilistic rankings that were calculated in section 3.5.4.2 for the 1,014 accident locations that are currently considered as most dangerous. These results showed that in comparison with the use of accident count data, the use of Bayesian estimation techniques caused a different ranking order for the accident sites in term of dangerousness. Accordingly, when investigating different subsets of the 1,014 dangerous accident sites, this will also involve different accident sites and can give different results.

In particular, table 5.3 shows for different subsets of the 1,014 accident locations from 1997-1999, based on the probabilistic rankings, the number of accident sites that were also dangerous in 1999-2001.

These results show that when analyzing the top 15% most dangerous accident sites of 1997-1999 (152 sites), based on the probabilistic rankings, 69.08% of these accident sites are still dangerous in 1999-2001. Closer investigation of these sites showed that 8 accident sites that were ranked in the top 10 of most dangerous accident locations in 1997-1999 based on their probabilistic ranking are also dangerous in the period 1999-2001. For the top 40%, top 70% and top 800 these figures diminish respectively to 59.36%, 50.98% and 49.25% accident sites that are dangerous in both periods.

Comparing these results from table (5.3)with the results from table (5.1) shows that using Bayesian estimation values to determine the subsets of most dangerous accident sites lightly decreases the number of accident locations that are considered as dangerous in both periods for the defined subset. In other words, the use of probabilistic rankings will alter the accident sites that are considered as belonging to the subset of most dangerous accident sites.

Table 5.3: Common dangerous accident sites based on Bayes estimates

1997-1999 Bayes	**1999-2001**
top 15% (152 sites)	69.08%(105 sites)
top 40% (406 sites)	59.36% (241 sites)
top 70% (710 sites)	50.98% (362 sites)
top 800 (800 sites)	49.25% (394 sites)

Table 5.4: New dangerous accident sites based on Bayes estimates

1999-2001	**1997-1999**
top 15% (149 sites)	26.17 % (39 sites)
top 40% (397 sites)	40.55% (161 sites)
top 70% (694 sites)	48.55% (337 sites)
top 800 (800 sites)	50.62% (405 sites)

Indeed, as explained in section 2.1.2, it is possible that the high crash rates observed at some sites in 1997-1999 may be due to chance, or a combination of both chance and a moderately hazardous nature. In that case, a dangerous accident site may be nothing more than a meaningless cluster of accidents which occurs randomly and has no intrinsic meaning. These sites are likely to have fewer crashes in 1999-2001 because the number of crashes will tend to gravitate towards the long-term mean value.

Conversely, table (5.4) shows for different subsets of the 991 dangerous accident locations from 1999-2001, based on the probabilistic rankings, the number of accident sites that were not belonging to the most dangerous accident sites in 1999-2001.

Results from table (5.4) show that from the top 15% most dangerous accident sites (based on the probabilistic rankings) in 1999-2001, 26.17% accident locations were not even considered as dangerous based on the data from 1997-1999. Analogously with the results from table (5.2), the larger the subset, the higher the number of accident sites that are not considered as dangerous based on the data from 1997-1999, but should be considered as dangerous sites based on the data from 1999-2001 (respectively 40.55%, 48.55% and 50.62% for the top 40%, top 70% and the top 800).

Again, comparing these results from table (5.4)with the results from table (5.2) shows that using Bayesian estimation values to determine the subsets of most dangerous accident sites lightly decreases the number of accident sites that are considered as dangerous in both periods for the defined subset.

5.2.4 Buffering Dangerous Locations

As explained in section 5.2.1, the identification of dangerous accident locations in Flanders in theory is related to intersections with a radius of 50 meters or roadway segments of numbered roads with a length of 100 meters. Since measuring and calibration techniques continuously improve between periods, comparing the exact locations of the dangerous accident sites in the two periods would result in an underestimation of the actual number of dangerous accident sites that correspond between the two periods. Therefore, we rounded down the location data of the dangerous accident sites on road segments for the two periods to 100 meters. This resulted in 470 dangerous accident sites that correspond between 1997-1999 and 1999-2001. We concluded that this indicated that 68.21% of the dangerous accident sites from 1999-2001 were not considered as dangerous in 1997-1999.

However, in order to investigate whether these accident sites should in fact be considered as completely 'new' hazardous locations, we should examine whether these dangerous sites are not the result of accident migration factors. As explained in section 2.1.5, a relation has been observed between a hazardous accident site disappearance and the development of a new dangerous accident location near the previous one. This phenomenon of accident migration occurs when a dangerous accident site is not managed properly: on the one hand the accident number on the existing hazardous site decreases, while, on the other hand, usually in a road section nearby, the accident number suddenly increases.

Therefore, in order to investigate whether the dangerous accident locations indeed tend to migrate over time in the region of Flanders, we used a buffering technique to analyze whether the 'new' hazardous accident locations in 1999-2001 are not just accident sites that migrated from one dangerous site (based on the accident data of 1997-1999) to a neighboring location.

More specifically, we used a Geographical Information System to create buffers around the 1,014 dangerous accident sites from 1997-1999. In particular, we used the exact locations of the dangerous accident sites for 1997-1999 to create the buffers around these pinpoint locations. Next, we calculated how many of the 991 dangerous accident sites from 1999-2001 lie within at least one of these buffers. Table 5.5 shows the results for this research, depending on the value we chose for the buffer radius.

Results from table 5.5 show that when considering a buffer of 50 meters around the 1,014 dangerous accident sites from 1997-1999, 538 accident locations from the 991 dangerous accident sites from 1999-2001 correspond with at least one of these buffers. Comparing this value of 538 with the 470 accident sites we calculated before shows that, although both calculations rounded down the exact locations of the dangerous accident sites up to 50 meters, the buffering technique aggregates all the locations that lie in a circle with a radius of 50 meters and not just the accident sites that are

Table 5.5: Buffering the dangerous accident sites from 1997-1999

radius	**1999-2001**
50 meters	54.28 % (538 sites)
100 meters	57.41 % (569 sites)
150 meters	60.14 % (596 sites)
200 meters	62.16 % (616 sites)
250 meters	63.16 % (626 sites)
300 meters	64.48 % (639 sites)
350 meters	65.99 % (654 sites)
400 meters	67.10 % (665 sites)
450 meters	68.42 % (678 sites)
500 meters	70.23 % (696 sites)
1,000 meters	79.72 % (790 sites)

segments of the same road. Obviously, the larger the radius that is considered for buffering, the more dangerous accident sites from 1999-2001 will fall within one of these buffers (up to 79.92% when constructing buffers with a radius of 1,000 meters).

Based on the figures from table 5.5, one could argue that it does seem that accident locations tend to migrate over time in Flanders. Indeed, the construction of buffers shows that by taking into account the neighboring accident sites when comparing the locations of the dangerous sites in 1997-1999 and 1999-2001, the number of accident sites that should be considered as dangerous in both periods increases.

More specifically, table 5.6 shows for the 521 'new' dangerous accident locations from 1999-2001 the number of sites that now fall within at least one of the buffers of dangerous sites for 1997-1999. These figures show that for a radius of 50 meters 13.6% of the dangerous sites that are 'new' in 1999-2001 actually lie within the neighborhood of 50 meters of an accident site that is considered as dangerous based on the data from 1997-1999. Similar with the results from table 5.5, the larger the radius that is considered for buffering, the more 'new' dangerous accident from 1999-2001 will fall within one of these buffers (up to 61.61% when constructing buffers with a radius of 1,000 meters).

However, one should be careful when interpreting these results. As explained before, the buffering technique aggregates all the locations that lie with a certain

Table 5.6: Accident migration based on buffers

radius	**1999-2001**
50 meters	13.6 % (71 sites)
100 meters	19.19 % (100 sites)
150 meters	24.37 % (127 sites)
200 meters	28.21 % (147 sites)
250 meters	30.13 % (157 sites)
300 meters	32.63 % (170 sites)
350 meters	35.51 % (185 sites)
400 meters	37.62 % (196 sites)
450 meters	40.12 % (209 sites)
500 meters	43.57 % (227 sites)
1,000 meters	61.61 % (321 sites)

radius of the exact location and not just the accident sites that are segments of the same road. This technique has the advantage that it allows to identify dangerous accident locations that migrated from one road segment to a neighboring road segment that is located on a different road. When studying the effect of accident migration, this could be a correct result when the neighboring road segment now has an increased number of accidents as a direct result of the management of the previous dangerous accident site. However, this is not always necessarily the case. Therefore, more profound research on the road functions, road directions, road characteristics and management measures of the concerning road segments is necessary before any definite conclusions on accident migration can be made.

5.3 Conclusions

In this research, we used the accident data from 1997-1999 and 1999-2001 of the region of Flanders to investigate whether the dangerous accident sites in these periods, based on the priority values, tend to migrate over time.

First of all, by mapping the dangerous locations up to 100 meters precision, (this corresponds with the definition of dangerous accident locations in Flanders), we found that 470 dangerous accident are considered as dangerous in both periods. This indicates that 46.36% of the dangerous accident locations in 1997-1999 should still be considered as dangerous based on the data from 1999-2001 while 47.43% of the dangerous accident sites from 1999-2001 were not considered as dangerous in 1997-1999. A closer investigation of the 470 accident sites that are common in the two periods shows that the ranking of these accident locations based on the data from 1997-1999 varies between 1 and 1,014.

Next, results showed that using probabilistic rankings based on Bayesian estimation values to determine the subsets of most dangerous accident sites lightly decreases the number of accident locations that are considered as dangerous in both periods for the defined subset. Indeed, it is possible that the high crash rates observed at some sites in 1997-1999 may be due to chance, or a combination of both chance and a moderately hazardous nature. In that case, a dangerous accident site may be nothing more than a meaningless cluster of accidents which occurs randomly and has no intrinsic meaning. These sites are likely to have fewer crashes in 1999-2001 because the number of crashes will tend to gravitate towards the long-term mean value.

Furthermore, in order to investigate whether the dangerous accident locations indeed tend to migrate over time in the region of Flanders, we used a buffering technique to analyze whether the 'new' hazardous accident locations in 1999-2001 are not just accident sites that migrated from one dangerous site (based on the accident data of 1997-1999) to a neighboring location. These results showed that by taking into account the neighboring accident sites when comparing the locations of the dangerous sites in 1997-1999 and 1999-2001, the number of accident sites that should be considered as dangerous in both periods increases. However, although this technique has the advantage that it allows to identify dangerous accident locations that that are located in the neighborhood of a different dangerous accident location, one should be careful when interpreting these results. More profound research on the road functions, road directions, road characteristics and management measures of the concerning road segments is necessary before any definite conclusions on accident migration can be made. Therefore, as explained before (see section 2.1.5) to prevent accident migration, government should be careful to determine the target accidents for a treatment. These target accidents are all those that can be affected by the treatment. It follows that a convincing evaluation of the safety effect of a treatment requires a good understanding of the process by which the accidents are generated and avoided.

Finally, note that in this research, due to limitations in the availability of the data, the data are not independent, as the year 1999 is included in both data sets

(1997-1999 and 1999-2001). Inclusion of the same year in both data sets is bound to create dependence between them, which reduces variance and makes the black spots appear more stable than they actually are. Secondly, this research was based on the counted accident data in the two periods. As explained before, this is very sensitive to random variation in accident counts and to the regression to the mean problem (see section 2.2.1). Further research is necessary in order to account for this regression to the mean problem and controlling spatial dependency of the accident counts.

Chapter 6

Final conclusions

In this chapter, an overview of the main research contributions of this dissertation is presented. More specifically, we will present the most important conclusions from this research and discuss the limitations and topics for future research. This chapter will mainly be based on the chapters 3, 4 and 5 of this dissertation.

6.1 Research Objectives

In general, typical procedures for hazardous site correction can be divided into three basic tasks (Schlüter (1997), Vistisen (2002)) :

1. The identification and ranking of hazardous locations. This results in a list of sites with promise (‘Identification Phase’).

2. Prioritizing these sites by diagnosing the problems at identified locations and determining potential remedial treatments in order to identify cost-effective safety improvement projects(‘Investigation Phase’).

3. The appraisal of alternative treatments followed by implementation of the best treatment if sufficiently cost-effective. To evaluate the effect of treatment, before and after studies need to be conducted. (‘Program Implementation Phase’)

In this dissertation, we focus on the first and second element of these steps, namely the identification and investigation of hazardous accident locations. Accordingly, in this research, we had two main research objectives.

First, we wanted to investigate how hazardous accident locations are currently identified and ranked in Flanders (the Flemish speaking community of Belgium). More specifically, the objective was to perform a sensitivity analysis, based on the same data used to select and rank the 1,014 accident sites that are currently considered as dangerous by the Flemish government, to investigate the strengths and weaknesses of this approach to rank and select dangerous accident sites. Second, in order to develop effective countermeasures to reduce the number of accidents at dangerous locations, one should properly and systematically relate accident frequency and severity to a large number of variables. Accordingly, the second objective of this research was to investigate how data mining and statistical techniques can be implemented to identify and profile dangerous accident locations in terms of accident related data and location characteristics. Additionally, we aimed to develop new models, using these techniques, which will provide new insights into the criteria that aim to explain to probability of the occurrence of a traffic accident. Finally, we were also interested in investigating whether the dangerous accident sites in Flanders tend to migrate over time using accident data from different time periods.

6.2 Ranking and Selecting Dangerous Accident Locations

In chapter 3 of this dissertation, we showed that changing the 1_ 3_ 5 weighing values that are currently used for respectively a lightly injured, seriously injured and deadly injured person will have an important impact on the selection of the most dangerous accident locations. However, the choice for the values of these weight parameters are mainly a policy decision depending on the government's priorities in the traffic safety policy. Therefore, no conclusions were made concerning which weighing criterion should be preferred. This was not the objective of this research and will require additional data and in depth analysis of the accident locations.

Furthermore, we explained that by giving weight to the severity of the accident we can correct for the bias in the priority score that occurs when the number of occupants of the vehicles are subject to coincidence. However, one should investigate whether the number of occupants, and accordingly the number of injured persons, is not a coincidence but more likely a trend. For these locations, correcting for the number of passengers would not be advisable since the number of injuries that appear at these locations are inherent to the locations characteristics.

Next, we introduced the use of Bayesian estimation values instead of historic count data to rank the accident locations in order to take into account the problem of random variation in accident counts. Additionally, we developed Bayesian ranking plots that can be used to visualize the estimated probability that a location will be ranked as belonging to the most dangerous locations. These probability plots can provide policy makers with a scientific instrument with intuitive appeal to select dangerous road locations on a statistically sound basis. In particular, using this tool, we showed that the choice for tackling only 800 accident locations, as the government currently does, seems somewhat arbitrary. However, no statement was made in this research on how many sites should be tackled since this, of course, will depend on the budget that is available to improve traffic safety.

As a conclusion, we presented the results of using a valuation of casualties based on direct costs, indirect costs and validation for human suffering to give weight to the accidents. This valuation resulted in the weighing values 1_ 7_ 33 when the most severe injury respectively concerns a light, serious or deadly injury. Based on estimates from a hierarchical Bayes model, we generated probability plots, in order to visualize the estimated probability that a location will be ranked as dangerous.

Results showed that combining these ranking criteria will have a big impact on the selection and ranking of dangerous accident locations. Therefore, considering this impact quantity, we want to sensitize government to carefully choose the criteria for ranking and selecting accident locations without stating that the criterion used in this paper should be preferred to the currently used ranking method. Indeed, the 1_ 7_ 33 weighing values seem a reasonable and scientifically sound alternative for the currently used 1_ 3_ 5 weighing combination, but it is up to the government to decide which priorities should be stressed in their traffic safety policy. Then, the according weighing value combination can be chosen to rank and select the most dangerous accident locations.

In this context, we developed a mixed integer programming model in order to automatically generate the 'optimal' weighing values for the 800 accident locations that are primarily responsible for the human suffering on the Belgian roads. These results showed that in order to select the accident locations that maximize the human suffering on the roads, the government should not only take into account the valuation of injury types from an economic or ethical point of view but also the relation between the number of injury types at each location.

As a final remark, we want to point out that, although this was the main focus of this dissertation, in practice one should not only rank the accident locations based on the benefits that can be achieved from tackling these locations. One should also incorporate the costs of infrastructure measures and other actions that these accident sites require in order to enhance the safety on these locations. By balancing these costs and benefits against each other, the accident locations can then be ranked according to the order in which they should be prioritized.

6.3 Profiling Hazardous Accident Locations

In chapter 4 of this dissertation, we used an association algorithm used on a data set of road accidents to profile hazardous accident locations in terms of accident related data and location characteristics. We showed that frequent item sets and association rules can be generated to identify accident circumstances that frequently occur together. More specifically, we used this technique to discern frequently occurring accident types, each with different relevant accident conditions. Furthermore, we were able to identify and profile frequently occurring accident patterns at high frequency accident locations and determine the degree in which these accident characteristics are discriminating between high frequency and low frequency accident locations. Additionally, we found that spatial concentrations of accidents are characterized by specific accident circumstances, which require different countermeasures to reduce their number such as improvements in terms of road design, signalization, and local environment. From this research, we concluded that a special traffic policy towards accidents inside black zones and accidents outside these zones should be considered.

Note however that the use of this technique is of an explorative character since it describes the co-occurrence of accident circumstances but it does not give any explanation about the causality of these accident patterns. Its role is to give direction to more profound research on the causes of these accident patterns and explanation which will require the use of some additional techniques or expert knowledge. These patterns represent interesting interactions in accident factors, which accordingly can be used to test in statistical models. Furthermore, the use of frequent item sets not only allows to give a descriptive analysis of accident patterns and discern different accident types, it also creates the possibility to find the accident characteristics to find the accident characteristics that are discriminating between groups such high frequency accident locations and low frequency accident locations or black zones and non black zones.

Although the analyses carried out in this research revealed several interesting patterns which, in turn, provide valuable input for purposive government traffic safety actions, we pointed out some limitations of this research. First, the skewed character of the accident data limits the amount of information contained in the data set and will therefore restrict the number of circumstances that will appear in the results. Moreover, the choice for the minimum support parameter can prevent the association algorithm from generating rules on the less frequent accident conditions. However, this information on rare accident types could be very useful since the circumstances in which these accidents occur, will probably not be trivial and more difficult to discern. Additionally, the inclusion of domain knowledge in the association algorithm would improve the mining capability of this data mining technique and would facilitate the post-processing of the association rules set to discover the most interesting accident patterns. Moreover, the variables used here are restricted to those collected by the police. Ideally, our range of variables should therefore be extended to other sources (traffic, land-use, accessibility, etc.) in order to better approach the explanation process. Finally, considering the large number of attributes in the traffic accident data set, we showed that, although it is possible to generate a large number of item sets with more than 4 items, for this research the increased complexity of these longer patterns is not compensated by the more explanatory power of these item sets.

6.4 Evolution of Dangerous Accident Sites over Time

Finally, in chapter 5 of this dissertation, we used the accident data from 1997-1999 and 1999-2001 of the region of Flanders to investigate whether the dangerous accident sites in these periods, based on the priority values, tend to migrate over time.

First of all, we found that 46.36% of the dangerous accident locations in 1997-1999 should still be considered as dangerous based on the data from 1999-2001 while 47.43% of the dangerous accident sites from 1999-2001 were not considered as dangerous in 1997-1999. A closer investigation of the 470 accident sites that are common in the two periods shows that the ranking of these accident locations based on the data from 1997-1999 varies between 1 and 1,014. Next, we illustrated, using probabilistic rankings based on Bayesian estimation values, that the high crash rates observed at some sites in 1997-1999 may be due to chance, or a combination of both chance and a moderately hazardous nature. These sites are likely to have fewer crashes in 1999-2001 because the number of crashes will tend to gravitate towards the long-term mean value. Finally, we used a buffering technique to analyze whether the 'new' hazardous accident locations in 1999-2001 are not just accident sites that migrated from one dangerous site (based on the accident data of 1997-1999) to a neighboring location. However, although this technique has the advantage that it allows to identify dangerous accident locations that that are located in the neighborhood of a different dangerous accident location, one should be careful when interpreting these results. More profound research on the road functions, road directions, road characteristics and management measures of the concerning road segments is necessary before any definite conclusions on accident migration can be made. Accordingly, to prevent accident migration, government should careful to determine the target accidents for a treatment. These target accidents are all those that can be affected by the treatment. It follows that a convincing evaluation of the safety effect of a treatment requires a good understanding of the process by which the accidents are generated and avoided.

Bibliography

[1] Abdel-aty, M. and A. Radwan (2000). Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32 (5), 633-642.

[2] Abdel-Aty, M. and H. Abdelwahab (2004). Modeling rear-end collisions including the role of drivers visibility and light truck vehicles using a nested logit structure. Accident Analysis and Prevention 36 (3), 447-456.

[3] Agent, K.,R. and R.C. Deen (1975). Relationship between roadway geometrics and accidents. Transportation Research Record 541, Washington DC.

[4] Agrawal, R., Imielinski, T. and A. Swami (1993). Mining association rules between sets of items in large databases. Proceedings of ACM SIGMOD Conference on Management of Data, Washington D.C., USA, May 26-28, 207-216.

[5] Agrawal, R., Mannila, H., Srikant, R., Toivonen R., Verkamo, H., (1996). Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, California, USA, 307-328.

[6] Ali, K., Manganaris, S. and R. Srikant (1997). Partial classification using association rules. Proceedings of the third international conference on knowledge discovery and data mining, Newport Beach, California (USA), 115-118.

[7] Ali, S. (2001). Pedestrianvehicle crashes and analytical techniques for stratified contingency tables. Accident Analysis and Prevention 34 (2), 205-214.

[8] Anand, S., Bell,D., Hughes, J. and A. Patrick (1997). Tackling the cross sales problem using data mining. Proceedings of the 1st International Conference On Knowledge Discovery and Data Mining, 331-343

[9] Andrey, J., Mills, B. and J. Vandermolen (2001). Weather information and road safety. Institute for Catastrophic Loss Reduction, Paper Series No. 15, Ontario, Canada.

[10] Anselin, L., (1995). Local indicators of spatial association-LISA. Geographical Analysis 27 (2), 93-115.

[11] Asian Development Bank (2003). Road safety guidelines for the Asian and Pacific Region. Guidelines for Decision Makers on Road Safety Policy. ISBN: 971-561-113-3.

[12] The Bureau of Transport and Regional Economics of Australia (2001), The Black Spot Program 1996-2002: An evaluation of the first three years.

[13] Baker S., Waller A. and J. Langlois (1991). Motor vehicle deaths in children: geographic variations. Accident Analysis and Prevention 23, 19-28.

[14] Belanger, C. (1994). Estimation of safety of four-legged unsignalized intersections. Transportation Research Record, Journal of the Transportation Research Board, No. 1467, TRB. National Research Council, Washington, DC, 23-29.

[15] Berry, M. and G. Linoff, (1997). Data Mining Techniques for Marketing, Sales, and Customer Support, John Wiley and Sons.

[16] Bloemer, J., Brijs, T., Swinnen, G. and K. Vanhoof (2002). Identifying latently dissatisfied customers and measures for dissatisfaction management. Journal of Bank Marketing 20 (2), 27-37.

[17] Blower, D., Campbell, K. and P. Green (1993). Accident rates for heavy truck-tractors in Michigan. Accident Analysis and Prevention 25 (3), 307-321.

[18] Boyle, A.J. and C.C. Wright (1984). Accident 'migration' after remedial treatment of accident blackspots. Traffic Engineering and Control, 25(5), 260-267.

[19] Braddock M., Lapidus G., Cromley E, Cromley R., Burke G and L. Banco (1994). Using a geographic information system to understand child pedestrian injury. American Journal of Public Health 84, 1158-1161.

[20] Brin, S., Motwani, R. and C. Silverstein (1997). Beyond market baskets: generalizing association rules to correlations. In Proceedings of the ACM SIGMOD Conference on Management of Data, Tucson, Arizona, USA, May 13-15, 265-276.

[21] Brin, S., Motwani, R. and C. Silverstein (1998). Beyond market baskets: generalizing association rules to dependence rules. In Data Mining and Knowledge Discover, 2 (1), 39-68.

[22] Brijs T. (2002). Retail Market Basket Analysis. A Quantitative modeling approach. PhD Dissertation. University of Hasselt, Diepenbeek, Belgium.

[23] Brijs T., Karlis D., Van den Bossche F. and G. Wets (2003). A Bayesian model for ranking hazardous sites. Proceedings of the 11th Symposium Statistical Software, the Royal Dutch Academy of Sciences, Amsterdam, The Netherlands, 55-72.

[24] Brodsky, H. and A. Hakker (1988). Risk of a road accident in rainy weather. Accident Analysis and Prevention, 20, 161-176.

[25] Cameron, M. (1997) Accident data analysis to develop target groups for countermeasures. Monash University Accident Research Centre, Reports 46 and 47.

[26] Cameron, A. and P. Trivedi (1986) Econometric models based on count data: comparisons and applications of some estimators and test. Journal of Applied Econometrics, 1 29-55.

[27] Casaer, F., Eckhardt, N., Steenberghen T., Thomas, I., Wets G. and J. Wijnant (2003). Een onderzoek naar de kwaliteit van de Belgische ongevallendata (in Dutch). Working paper, University of Hasselt, Diepenbeek.

[28] Ceder, A. and M. Livneh (1982). Relationships between road accidents and hourly trafficflow. Accident Analysis and Prevention 14 (1), 19-34.

[29] Chen, W. and P. Jovanis (2002). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record 1717, Washington, D.C., 1-9.

[30] Cheng W. and S. Washington (2005). Experimental Evaluation of Hotspot Identification Methods. Accident Analysis and Prevention 37 (5), 870-881.

[31] Congdon P. (2003). Applied Bayesian Modelling. John Wiley and sons Ltd, England.

[32] Cox, D. (1983).Some remarks on overdispersion. Biometrica, 70, 269-274.

[33] Christiansen, C., Moris, C. and O. Pendleton (1992). A hierarchical Poisson model with beta adjustments for traffic accident analyses. Center for statistical Sciences Technical Report 103, University of Texas, Austin.

[34] Danils, S. (2005). Verkeersonveiligheid in Vlaanderen: berusten of topprioriteit (in Dutch). Year Book Traffic Safety 2005, 5-6

[35] Dasgupta, A. and D. Pearce (1972), Cost-benefit analysis, theory and practice. Macmillian, England.

[36] Davies, J. (1990). A Bayesian analysis of some accident data. Statistician, 39, 11-17.

[37] Davis, C. (1986). Regression to the mean. Encyclopedia of statistical sciences, 7, 706-708. John Wiley and Sons, Inc.

[38] Davis, G.A. and S. Yang (2001). Bayesian identification of high-risk intersections for older drivers via Gibbs sampling. Transportation Research Record, 1746, TRB, National Research Council, Washington, D.C., 84-89.

[39] De Brabander B. and L. Vereeck (2005). Verkeersongevallen in België kosten jaarlijks 12,5 miljard euro (in dutch). Verkeersspecialist 122, 23-26 .

[40] De Groote, P. and V. Truwant (2003). Demografie en Samenleving (in Dutch). Leuven: Universitaire Pers Leuven.

[41] De Keersmaecker M., Frankhauser P. and I. Thomas I (2004) Analyse de la ralit fractale priurbaine : l'exemple de Bruxelles (in French). L'Espace Gographique, 2004 (3), 219-240.

[42] Doherty, S., Andrey, J. and C. MacGregor (1998). The situational risks of young drivers: the influence of passengers, time of day, and day of week on accident rates. Accident Analysis and Prevention 30(1), 45-52.

[43] Eckhardt N., Flahaut B. and I. Thomas (2004) Spatio-temporalit des accidents de la route en priphrie urbaine. L'exemple de bruxelles (in French). Recherche Transports et Scurit. 82, 35-46.

[44] Edwards J. (1996). Weather-related road accidents in England and Wales: a spatial analysis. Accident Analysis and Prevention 4 (3) 201-212.

[45] Elvik, R. (1997). Evaluations of road accident blackspot treatment: a case of the iron law of evaluation studies? Accident Analysis and Prevention, 29(2), 191-199.

[46] Elvik, R. and T. Vaa (2004). The handbook of road safety measures, Elsevier.

[47] Elvik, R. (2006). A new approach to accident analysis at hazardous road locations. TRB paper 06-0233. Forthcoming in Transportation Research Record.

[48] European Union Road Federation (2002). Good-practice guidelines to infrastructural road safety. October 2002.

[49] Fayyad, U., Piatetsky-Shapiro, G. and P. Smyth (1996). From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press / The MIT Press, 1-34.

[50] Flahaut B. and I. Thomas (2002). Identifier les zones noires d'un rseau routier par l'autocorrlation spatiale locale (in French). Revue Internationale de Gomatique 12 (2), 245-261.

[51] Flahaut, B., Mouchart, M., San Martin, E. and I. Thomas (2003). The local spatial autocorrelation and the kernel method for identifying 'black' zones. A comparative approach. Accident Analysis and Prevention 35 (6), 991-1004.

[52] Flahaut, B. (2004-a). Impact of infrastructure and local environment on road insecurity. Logistic modeling with spatial autocorrelation. Accident Analysis and Prevention. 36, 1055-1066.

[53] Flahaut, B. (2004-b). Towards a sustainable road safety in Belgium. Location of spatial concentrations of road accidents and explanatory modelling. PhD dissertation, Department of Geography, Louvain-la-Neuve, February 2004, 107 pp.

[54] Foldvary, L. (1979). Road accident involvement per miles travelled. Accident Analysis and Prevention 11, 75-99.

[55] Frawley, W., Piatetsky-Shapiro, G., and C. Matheus (1991). Knowledge discovery in databases: an overview. Knowledge Discovery in Databases. AAAI Press/ MIT Press, Menlo Park, California, USA, 1-27.

[56] Friedman, J. (1997). Data mining and statistics: What's the connection? Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics. http://wwwstat. stanford.edu/ jhf/ftp/dm-stats.ps

[57] Gelman, A., Carlin, J., Stern, H. and D. Rubin (1995). Bayesian Data Analysis. Chapman and Hall, London.

[58] Geurts K., Wets G., Brijs T. and Vanhoof K. (2002), The use of rule based knowledge discovery techniques to profile black Spots. The 6th Design and Decision Support Systems in Architecture and Urban Planning Conference, Ellecom, The Netherlands, July 7-10.

[59] Geurts, K. and Wets, G. (2003). Black Spot Analysis Methods: Literature Review. Rapport Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit, RA-2003-07, Diepenbeek.

[60] Geurts, K. Wets, G., Brijs T. and K. Vanhoof (2003). Profiling high frequency accident locations using association rules. Journal of Transportation Research Board, 1840, 123-130. Also in Electronic Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington, January 12-16, USA, pp. 18.

[61] Geurts K., Wets G., Brijs T. and K. Vanhoof (2003). Clustering and profiling traffic roads by means of accident data. Electronic proceedings of the European Transport Conference, Strasbourg, France, October 8-10.

[62] Geurts, K., Wets. G., Brijs, T. and Vanhoof, K. (2004), Identification and Ranking of Black Spots: Sensitivity Analysis. Journal of Transportation Research Board, 1897, 34- 42. Also in Electronic Proceedings of the 83th Annual Meeting of the Transportation Research Board, Washington, January 11-15, USA, pp. 17.

[63] Geurts K. (2004). Grote en kleine middelen om de verkeersveiligheid te verhogen: Hoe rangschikken en selecteren we gevaarlijke punten (in Dutch)? Jaarboek Verkeersveiligheid 2004, Vlaams Congres Verkeersveiligheid, 2004, Brussels, 44-46.

[64] Geurts K., Wets G., Brijs T. and K. Vanhoof (2004). Identifying and ranking dangerous accident locations: Overview Sensitivity analysis. Forthcoming in Proceedings of 17th ICTCT Workshop in Tartu, Estonia.

[65] Geurts K. (2005). Selectie en rangschikking van gevaarlijke punten: een grafische benadering (in Dutch). Jaarboek Verkeersveiligheid 2005, Vlaams Congres Verkeersveiligheid, 2005, Brussel.

[66] Geurts K., Wets G., Brijs T. and K. Vanhoof K. (2005). Ranking and selecting dangerous accident locations: Case Study. Urban Transport XI, eds. Brebbia and Wadhwa, WIT Press, 229-238.

[67] Geurts, K., Thomas I. and G. Wets (2005). Understanding accidents in black zones using frequent itemsets. Accident analysis and prevention 37 (4), 787-799.

[68] Geurts K., Wets G., Brijs T. Karlis, D. and K. Vanhoof (2006). Ranking and selecting dangerous accident locations: correcting for the number of passengers and bayesian ranking plots. Journal of Safety Research 37, 83-91.

[69] Goldstein H. and D. Spiegelhalter (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). In Journal of the Royal Statistical Society A 159, 385-443.

[70] Grahne G. and J. Zhu (2003). Efficiently using prefix-trees in mining frequent itemsets. Workshop on Frequent Itemset Mining Implementations. Melbourne, FL.

[71] Greibe, P. (2003). Accident prediction models for urban roads. Accident Analysis and Prevention 35, 273-285.

[72] Gupta, S.S. and J.C. Hsu (1980). Subset selection procedures with application to motor vehicle fatality data in a two-way latout. Technometrics, 22(4), 543-546.

[73] Hamaoka, H., Nagashima, H. and S. Morichi (1999). An analysis of the cause of traffic accidents at the black spots. Selected Proceedings of the 8th World Conference on Transport Research (2).

[74] Hammitt, J.K. (2002). QUALYs versus WTP. Risk Analysis 22 (2).

[75] Harwood, D., Council F., Hauer E., Hughes W. and A. Vogt (2000). Prediction of the expected safety performance of the rural two-lane highways. Midwest Research Institute, FHWA-RD-99-207, Kansas City, Missouri.

[76] Hauer, E. and B. Persaud (1984). Problem of identifying hazardous locations using accident data. Transportation Research Record, 975, TRB, National Research Council, Washington, D.C., 36-43.

[77] Hauer, E. (1986). On the estimation of the expected number of accidents. Accident Analysis and Prevention 18 (1), 1986, 11-12.

[78] Hauer, E. and B.N. Persaud (1987). How to estimate the safety of rail-highway grade crossing and the effects of warning devices. Transportation Research Record 1114, 131-140.

[79] Hauer, E. and A. Hakkert (1988). Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1-11.

[80] Hauer, E. (1996). Identification of sites with promise. Transportation Research Record:Journal of the Transportation Research Board, No. 1542, TRB. National Research Council, Washington, DC, 54-60.

[81] Hauer, E. (1997). Observational before-after studies in road safety. Elsevier Science Ltd, Oxford.

[82] Hauer, E., Kononov J., Allery B. and M. Griffith (2002). Screening the road network for sites with promise. Report submitted to the Transportation Research Board to be considered for publication in 2002.

[83] Hauer, E., Allery B., Kononov J., and M. Griffith (2004). How best to rank sites with promise. Journal of the Transportation Research Board, 1897, TRB, National Research Council, Washington, D.C., 48-54.

[84] Heydecker, B. and J. Wu (1993). A knowledge-based system for road accident remedial work. Computing Systems in Engineering. 4 (2-3), 337-348.

[85] Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in empirical Bayes estimation. Accident Analysis and Prevention, 33 (6), 799-808.

[86] Higle, J. and J. Witowski (1988). Bayesian identification of hazardous locations (with discussion). Transportation Research Record, 1185, 24-36.

[87] Hipp, J., Gntzer U. and G. Nakhaeizadeh G. (2000). Algorithms for association rule mining- A general survey and comparison. SIGKDD Explorations 2(1), pp.58.

[88] Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning, Vol. 11, 1993, 63-90.

[89] Hosking, J., Pednault, E. and M. Sudan (1997). A statistical perspective on data mining. Future Generation Computer Systems 13, 117-134.

[90] Jackman, S. (2003). Bayesian modelling in the social sciences: an introduction to Markoc Chain Monte-Carlo. Department of Political Science, Stanford University, Stanford.

[91] Joly, M., Bourbeau, R., Bergeron, J., and S. Messier (1992). Analytical approach to the identification of hazardous road locations: a review of the literature. Centre de recherche sur les transports, Universit de Montral.

[92] Jovanis, P. and H. Chang (1989). Disaggregate model of highway accident occurrence using survival theory. Accident Analysis and Prevention 21 (5), 445-458.

[93] Julien A. and J. Carr (2002). Cheminements pitonniers et exposition au risque (in French). Recherche Transports Scurit 76, 173-189.

[94] Karlis, D. (2000) An EM Algorithm for multivariate Poisson distribution and related Models. Department of Statistics, Athens University of Economics and Business, Greece.

[95] Kavsek, B., Lavrac N. and J. Bullas (2002). Rule induction for subgroup discovery: A case study in mining UK traffic accident data. Proceedings of Conference on Data Mining and Warehouses (SiKDD2002), Ljubljana, Slovenia, October 15.

[96] Keall, M., Frith W. and T. Patterson (2005). The contribution of alcohol to night time crash risk and other risks of night driving. Accident Analysis and Prevention 37 (5), 816-824.

[97] Kim, D., Jutaek, O. and S. Washington (2006). Modeling crash outcomes: New insights into the effects of covariates on crashes at rural intersections. ASCE Journal of Transportation Engineering (in press).

[98] Kononov, J. and B. Janson (2002). Diagnostic Methodology for the detection of safety problems at intersections. Proceedings of the Transportation Research Board (CD-ROM), Washington D.C., USA, January 13-17. also in Tranpsortation Research Record 1794.

[99] Kulmala R. (1995). Safety at rural three- and four-arm junctions. Technical Research Centre of Finland (VTT), Espoo, Finland.

[100] Lammar P.(2003). Haalbaarheidsstudie voor correctie ongevallengegevens: tussentijds rapport (in Dutch). Policy Research Center for Traffic Safety, RA-2003-15, Diepenbeek.

[101] Land, K., McCall, P. and D. Nagin (1996). A comparison of Poisson, negative binmoial, and semi-paramteric mixed Poisson regression models-with empirical applications to criminal careers data. Sociological Methods and Research, 24 (4), 387-442.

[102] Larsen, L. and P. Kline (2002). Multidisciplinary in-depth investigations of head-on and left-turn road collisions. Accident Analysis and Prevention 34, 367-380.

[103] LaScala, E., Gerber D and P. Gruenewald P. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis and Prevention 32, 651-658.

[104] LaScala, E., Gruenewald, P. and F. Johnson (2004). An ecological study of the locations of schools and child pedestrian injury collisions. Accident Analysis and Prevention 36 (4), 569-576.

[105] Lee J. and F. Mannering (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34, 149-161.

[106] Lee, C., Saccomanno, F. and B. Hellinga (2002). Analysis of crash precursors on instrumented freeways. Proceedings of the Transportation Research Board (CD-ROM), Washington D.C., USA, January 13-17.

[107] Lee, C. and M. Abdel-Aty (2005). Comprehensive analysis of vehiclepedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37 (4), 775-786.

[108] Levine N. (2002). CrimeStat II : A spatial statistics program for the analysis of crime incident locations (version 2.0). Ned Levine and Associates: Houston, TX/National Institue of Justice : Washington, DC.

[109] Lindenbergh, S.D. (1998). Smartengeld (in Dutch). Ph.D. Dissertation, Leiden University, The Netherlands.

[110] Liu B., W. Hsu and Y. Ma. (1999). Pruning and summarizing the discovered associations. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 125-134 .

[111] Ljubic P., Todorovski L., Lavrac N. and J. Bullas (2002). Time series analysis of UK traffic accident data. Proceedings of Conference on Data Mining and Warehouses (SiKDD 2002), Ljubljana, Slovenia, October 15.

[112] Maher, M. (1990). A bivariate negative bimoial model to explain traffic accident migration. Accident Analysis and Prevention, 22(5), 487-498.

[113] Maher, MJ and I. Summersgill (1996). A Comprehensive Methodology for the Fitting of Predictive Accident Models. Accident Analysis and Prevention 28 (3), 281-296.

[114] Mannila, H. (1997). Methods and problems in data mining. Proceedings of the International Conference on Database Theory, Delphi, Greece, January 8-10, 41-45.

[115] Martin, J. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. Accident Analysis and Prevention, 34, 619-629.

[116] Mannila H. (2000). Theoretical Frameworks for Data Mining. SIGKDD Explorations 1 (2), 30-32.

[117] Maycock, G. and R.D. Hall (1994). Accidents at 4-arm roundabouts. Report 1120. Crowthorne, U.K., Transport and Road Research Laboratory.

[118] McCullagh, P. and J. Nelder (1989). Generalized linear models, 2nd edition, Chapman and Hall, London.

[119] McGuigan, D. (1981). The use of relationships between road accidents and traffic flow in black-spot identification. Traffic Engineering and Control, 22 (8-9), 448-453.

[120] McLachlan, G., and D. Peel (2000). Finite Mixture Models. Wiley Publications NY.

[121] Melchers, R. (2001). On the ALARP approach to risk management. Reliability Engineering and System Safety, 71 (2), 201-208.

[122] Merenne B., Van der Haegen H. and E. Van Hecke (1997). La Belgique. Diversit territoriale (in French). Bulletin du Crdit Communal, n202. Also available on Internet : http://www.belspo.be

[123] Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26 (4), pp.471- 482.

[124] Ministerie van Verkeer en Waterstaat (2001). Kerncijfers verkeersonveiligheid 2002 (in Dutch). http://www.rws-avv.nl. Accessed in November 2005).

[125] Ministry of the Flemish Community (2001). Ontwerp-Mobiliteitsplan Vlaanderen(in Dutch),Brussels, Belgium, http://viwc.lin.vlaanderen.be/mobiliteit. Accessed July 2003.

[126] Miranda-Morena, L., Fu L., Saccomanno F. and A. Labbe (2005). Alternative risk models for ranking locations for safety improvement. Proceedings of 84th Annual Meeting of the Transportation Research Board.

[127] Moran, P. (1948). The interpretation of statistical maps. Journal of the Royal Statistical Society 10b, 243-251.

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

[129] Nakahara S. and S. Wakai (2001). Underreporting of traffic injureis involving children in Japan. Injury Prevention (7), 242- 244.

[130] Nassar, S. (1996). Integrated Road Accident Risk Model. Phd. Dissertation, Waterloo, Ontario, Canada.

[131] Ng, K., Hung, W. and W. Wong (2002). An algorithm for assessing the risk of traffic accident. Journal of Safety Research (33), 387-410.

[132] Nguyen, T. (1991). Identification of accident 'black'spot locations, an overview. VIC Roads Safety Division, Research and Development Department, Australia.

[133] National Institute for Statistics (2002). FOD Economics - Department of Statistics, Statistics on road accidents. http://statbel.fgov.be/figures/download_nl.asp

[134] National Institute for Statistics and Belgian Institute for Traffic Safety (2001). Jaarrapport Verkeersveiligheid 2001 (in Dutch)(CD-ROM), BIVV v.z.w., Brussels.

[135] NCHRP Synthesis 336 (2004). National Cooperative Highway Research Program: Road Safey audits. A synthesis of highway practice. Transportation Research Board, Washington, D.C.

[136] Nord, E. (1999). Cost-value analysis in health care. Making sense of QUALYs. Cambridge, Cambridge University Press.

[137] Ogden, K.W.(1996). Safer Roads: a guide to road safety engineering. Published by Ashgate publishing limited, England.

[138] Oppe, S. (1979). The use of multiplicative models for analysis of raod safety data. Accident Analysis and Prevention 11, 101-115.

[139] Pact van Vilvoorde (2001) (in Dutch). Sociaal Economische Raad van Vlaanderen (in Dutch), www.serv.be (last accessed in November 2005).

[140] Persaud, B. (1990). Blackspot identification and treatment evaluation. The research and development branch, Ontario Ministry of Transportation.

[141] Persaud, B. and A. Kazakov (1994). A procedure for allocating a safety improvement budget among treatment sites. Accident Analysis and Prevention, 26 (1), 121-126.

[142] Persaud, B.N., C. Lyon and T. Nguygen (1999). Empirical Bayes procedure for ranking sites for safety investigation by potential safety improvements. Journal of the Transportation Research Board. Transportation Research Record 1665, 7-12.

[143] Petersen, J. H., Andersen, P. K. and Gill, R. D. (1996). Variance components models for survival data, Statistica Neerlandica 50(1), 193211.

[144] Pyle, D. (1999). Data preparation for data mining. San Francisko, CA: Morgan Kaufman.

[145] Rajalin S. (1994). The connection between risky driving and involvement in fatal accidents. Accident Analysis and Prevention 26 (5), 555-562.

[146] Saccomanno, F.F. and C. Buyco (1988). Generalized loglinear models of truck accident rates. Transportation Research Record 1172, 23-31.

[147] Saccomanno, F.F., Grossi, R. Greco D. and A. Mehmood (2001). Identifying black spots along highway SS107 in Southern Italy using two models. Journal of Transportation Engineering. American Society of Civil engineering, November.

[148] Schlüter, P., Deely, J.J and A. Nicholson (1997). Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. The Statistician 46, 293-316.

[149] Schwarz, G. (1978) Estimating the dimensions of a model. The Annals of Statistics 6, 461-464.

[150] Shankar, V., Mannering, F. and W. Barfield (1995). Effect of roadway and evironmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27 (3), 371-389.

[151] Silcock, D. and A. Smyth (1985). Methods of identifying accidents black spots. Transport Operations Research Group, Department of Civil Engineering, University Of Newcastle Upon Tyne.

[152] Smyth, P. (2001). Data mining at the interface of computer science and statistics. Data Mining for Scientific and Engineering Applications, eds. R. Grossman, C. Kamath, V. Kumar, Kluwer Academic Publishers, 35-63.

[153] Spiegelhalter, D., Thomas, A., Best, N. and W. Gilks (1996). Bugs 0.5: Bayesian inference using Gibbs sampling, manual (version ii). MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

[154] Spiehelhalter, D. (1999). Bayesian statistical analysis. AI and statistics. Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

[155] Staten-Generaal van de Verkeersveiligheid (2001) (in Dutch). Verslag van het begeleidingscomit aan het bestuurscomit, www.bivv.be (last accessed in November 2005).

[156] Strnad M., Jovic F., Vorko A., Kovacic L. and D. Toth (1998). Young children injury analysis by the classification entropy method. Accident Analysis and Prevention 30 (5), 689-695.

[157] Taber, J. (1998). Multi-objective optimization of intersection and roadway design. Utah Transportation Center, Utah State University.

[158] Tan, P, Kumar, V. and J. Srivastava (2002). Selecting the right interestingness measure for association patterns. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[159] Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis and Prevention 28, 251-264.

[160] Tingvall C. and N. Haworth (1999). Vision Zero: An ethical approach to safety and mobility. Paper presented to the 6th ITE International Conference Road Safety and Traffic Enforcement: Beyond 2000, Monash University Accident Research Centre, Melbourne, .

[161] Thomas, P., Morris, A., Otte, D. and J. Breen (2003). Real-world accident data - Coordinated methodologies for data collection to improve vehicle and road safety. 18th International Technical Conference on the Enhanced Safety of Vehicles, Japan.

[162] Tunaru, R. (1999), Hierarchical Bayesian models for road accident data. Traffic Engineering and Control, 40 (6), 318-324.

[163] Tunaru, R. (2002). Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31, 221-229.

[164] Valent F., Schiava, F., Savonitto C., Gallo, T., Brusaferro, S. and F. Barbone (2002). Risk factors for fatal road traffic accidents in Udine, Italy. Accident Analysis and Prevention 34, 71-84.

[165] Van den Bossche, F., Wets G. and E. Lesaffre (2002). A Bayesian hierarchical approach to model the rank of hazardous intersections for bicyclists using the Gibbs sampler. Flemish Research Center for Traffic Safety, Diepenbeek, Belgium, RA-2002-03.

[166] Vandersmissen, M., Pouliot, M. and D. Morin (1996). Comment estimer l'inscurit d'un site d'accident: tat de la question (in French). Recherche Transports Scurit 51, 49-60.

[167] Vistisen, D. (2002). Models and methods for hot spot safety work, PhD thesis, Denmark.

[168] Viveros, M., Nearhos, J., and J. Rothman (1996). Applying Data Mining Techniques to a Health Insurance Information System. Proceedings of the 22nd VLDB Conference, Bombay (India), 286-294.

[169] Vogelesang, A. (1996). Bayesian Methods in Road Safety Research: an Overview. Institute for Road Safety Research (SWOV), Leidschendam, The Netherlands, pp.44.

[170] Wedel, M., Desarbo, W., Bult, J. and V. Ramaswamy (1993). A latent class poisson regression model for heterogeneous count data. Journal of Applied Econometrics, 8, 397-411.

[171] Wilde, G. (1986). Beyond the concept of risk homeostasis: suggestions for research and application towards the prevention of accidents and life-style related disease. Accident Analysis and Prevention 18 (5), 377-401.

[172] Wong Y. and A. Nicholson (1992). Driver behaviour at horizontal curves: risk compensation and the margin of safety. Accident Analysis and Prevention 24 (4), 425-436.

[173] Yang, R. and J. Berger (1997). A catalog of Noninformative Priors. ISDS Discussion Paper 97-42, Department of Statistics, Purdue University.

[174] Yearbook Traffic Safety 2003. Vlaamse Stichting Verkeerskunde. (in dutch)

[175] Chelghoum N. and K. Zeitouni (2004). Spatial data mining implementation. Alternatives and performances. Presented at GeoInfo2004, Brazilian Imopsium on GeoInformatics, 22/11-23/11 2004.

Appendix1

List of variables included in table ONGEVAL.

Variable	Description	Attribute Values
Volgnummer	Unique accident ID	numeric
Prov	Province	Numeric
Jaar	Year	Numeric
RP	Police unit	Text
Eenheid	Police code	Numeric
Pvnr	Police report number	Numeric
Nis	City code	Numeric
Tijdstip	Hour, day, month, year	Numeric
Kruispunt	Intersection	1=near intersection, 2=outside intersection
In bbkom	Built-up area	1=inside built-up area, 2=outside built-up area
Kprgl1,2	Intersection traffic regulation	1=police officer, 2=traffic lights, 3=flashing light, 4=traffic signs, 5=priority to the right
Licht	Light conditions	1=daylight, 2=twilight, 3=public lighting, 4=night
Plakar1,2	Local characteristics	1=road works, 2=bridge, 3=tunnel, 4=railroad, 5=roundabout
Staatw1,2	Road surface characteristics	1=dry, 2=wet, 3=snow, 4=clean, 5=dirty, 9=unknown

Factorenw1,2	Road factors	1=bad road surface, 2=faulty signals, 3=faulty lighting, 4=road works, 5=queue, 6=downhill, 7=curve, 8=bad visibility
Varia1,2,3,4	Varia	1...14 *For example*: 1=accident, 2=aquaplaning, 3=sun-blinded, 4=school, 6=bus stop, 8=no seat belt, 9=no helmet, 10=no child seat
Weer1,2	Weather conditions	1=normal, 2=rain, 3=fog, 4=wind, 5=snow, 6=hail, 7=other, 9=unknown
Aantweggebrs	Number of road users	Numeric
Aantpass	Number of dead or injured passengers	Numeric
Aantslachtoffers	Victims (no road users)	Numeric
Totaal betrokken	Number of persons involved	Numeric
Totaaldoden	Number of deaths	Numeric
Totaalligew	Number of lightly injured	Numeric
Totaalzwgew	Number of seriously injured	Numeric
Dodenlg	Number of deaths due to light injuries	Numeric
Dodenzg	Number of deaths due to serious injuries	Numeric

Appendix2

List of variables included in table WEGGEBRUIKERS.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Gebruiker	Road user	Numeric
Locid	Unique location ID	Numeric
Zinverplaatsing	Direction	1=positive, 2=negative, 3=transverse, 4=not applicable, 5=unknown
Aard	Type of road user	1,...24 *For example*: 1=car, 6=truck, 10=bus, 13=motorbike, 14=moped, 18=bike, 20=pedestrian
Aantpass	Number of passengers	Numeric
Alcohol	Alcohol test	1=no, 2=refused, 3=positive, 4=negative
Geslacht	Gender	1=male, 2=female
Gevolgen	Consequences	1=dead, 2=seriously injured, 3 lightly injured, 4=uninjured
Land	Country of vehicle registration	Text
Leeftijd	Age	Numeric
Nrplaat	License plate	Text
Toest1,2	Condition	1=normal, 2=drunk, 3drugs, 4=ill

Typeaanrijding	Type of collision	1=multiple, 2=frontal, 3=parallel, 4=sideways, 5=pedestrian, 6=obstacle on roadway, 7=obstacle outside roadway, 8=one driver, no obstacle, 9=other, unknown
Tegenhindernis	Obstacle type	50...67 *For example*: 50=animal, 51=train, 52=tram, 59=tree
Tegenweggebr	Number of collided road user	Numeric
Beweging	Movement	1=straight direction, 2=opposite direction, 3=loss control to the left 4=loss control to the right, 5=left turn, 6=right turn, 7=pass left, 8=pass right, 9=u-turn, 10=drive backwards, 11=car breakdown, 12=standstill opening door, 13=standstill, 14=parking, 15=private property, 16=other movement
Dynamica	Dynamics	1=constand speed, 2=brake, 3=accelerate, 4=standstill, 5=unknown
Factorengebr1,2	Factors road user	1=ignored red light, 2=no priority, 3=over white line, 4=incorrect passing, 5=sidestep maneuver, 6=incorrect position on roadway, 7=loss control steering wheel, 8=no distance, 9=fall

Factorenv1,2	Factors vehicle	1=incorrect lighting, 2=bad tires, 3=broken tire, 4=defect trailer/cargo

Appendix3

List of variables included in table PASSAGIERS.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Gebruiker	Road user	Numeric
Passagier	Passenger	Numeric
Geslacht	Gender	1=male, 2=female
Leeftijd	Age	Numeric
Gevolgen	Consequences	1=dead, 2=seriously injured, 3=lightly injured
Plaats	Position in the vehicle	1=front, 2=back, 3=unknown

Appendix4

List of variables included in table VOETGANGERS.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Gebruiker	Road user	Numeric
Afstand	Unsheltered walking distance	Numeric
Overst	Visibility to other road users	1=visble, 2=not visible, 9=unknown
Plaats	Position	10=footpath, 11=cycle track, 20=out of vehicle, 30=right side roadway, 31=left side roadway, 40=zebra crossing with traffic lights, 41=zebra crossing with police officer, 42=zebra crossing, 43=next to zebra crossing with traffic lights, 44=next to zebra crossing with police officer, 45=next to zebra crossing, 46=no zebra crossing in 30m, 50=pedestrian not moving on roadway, 99=unknown

Appendix5

List of variables included in table FIETSERS.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Gebruiker	Road user	Numeric
Plaats	Position	1=separated cycle track, 2=marked cycle track on roadway, 3=other
Fietspad	Cycle track path	1=one way, 2=two way driving in normal direction, 3=two way driving in opposite direction

Appendix6

List of variables included in table SLACHTOFFERS.

Variable	Description	Attribute Values
Volgnummer	Unique accident ID	Numeric
Slachtoffer	Victim	Numeric
Geslacht	Gender	1=male, 2=female
Leeftijd	Age	Numeric
Gevolgen	Consequences	1=dead, 2=seriously injured, 3=lightly injured

Appendix7

List of variables included in table LOCATIE.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Locid	Unique Location ID	Numeric
Type	Type of numbered road	1=motorway, 2=district or province road
Ident	Unique road ID	Letter (A,B,N,R,P,T) + number
Wegtype	Type of numbered road	Letter (A,B,N,R,P,T)
Wegindex	Road index	Numeric
Kmp	Kilometer mark	Numeric
Nrgebouw	House number	Numeric
Snelheid	Maximum speed allowed	Numeric
Soort	Type of numbered road	1=one road lane, 2=separated road lanes
Straatnaam	Street name	Text
Straattype	Type of non-numbered road	Text

Appendix8

List of variables included in table LOCONG.

Variable	Description	Attribute Values
Volgnummer	Unique accident ID	Numeric
Locatie1,2	Unique location ID	Numeric

Appendix9

List of variables included in table GEVAARLIJKE PROD.

Variable	**Description**	**Attribute Values**
Volgnummer	Unique accident ID	Numeric
Gebruiker	Road user	Numeric
Borden	Signs	1=blank orange sign, 2=numbered sign
Opschrift1,2	Numbers on sign	Numeric
Staatlading	Condition of the load	1=empty, 2=leakage, 3=noleakage

Appendix10

List of items included in the analysis.

Variable	Item
Built-up area	inside built up area; outside built-up area
Type of road	highway; district or province road
Type of road lanes	road with one road lane; road with separated road lanes
Intersection	near intersection; outside intersection
Intersection traffic regulation	intersection police officer; signalized intersection; signalized intersection flashing light; intersection traffic signs; intersection priority to right
Location characteristics	road works; bridge; tunnel; railroad; roundabout
Road factors	bad road surface; faulty signals; faulty lighting; road works; queue; downhill; curve; bad visibility
Miscellaneous	accident following accident; aquaplaning; sun blinded; school; recreation center; bus stop; person swung out of vehicle; no safety belt; no helmet; no child seat; cargo on roadway before accident; cargo on roadway because of accident; fire after accident; comments
Weather conditions	normal weather; rain; fog; wind; snow; hail; other weather
Road conditions:	road surface : dry; wet; snow; clean; dirty
Light conditions	daylight; twilight; public lighting; night
Week	week; weekend
Day of week	Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday

Part of the day	morning rush hour (7-9 am); morning (10-12 am); afternoon (1pm-3pm); evening rush hour (4-6pm); evening (7-9pm); night (10pm-6am)
Type of road user	car; car double use; minibus; light truck; camper; truck; truck and trailer; truck; tractor; bus; trolley bus; motor coach; motorbike under 400cc; motorbike over 400cc; moped A; moped B; moped 3-4 wheels; bike; span; wheel chair; pedestrian with bike; pedestrian; horseman; other road user
Direction	positive way; negative way; transverse way; way not applicable
Movement	straight direction; opposite direction; loss control to the left; loss control to the right; left turn; right turn; pass left; pass right; u-turn; drive backwards; car breakdown; standstill opening door; standstill; parking; private property; other movement
Dynamics	constant speed; brake; accelerate; standstill
Alcohol	no alcohol test; refused alcohol test; positive alcohol test; negative alcohol test
Gender road user	male road user; female road user
Consequences road user	dead; seriously injured; lightly injured; uninjured
Age road user	0-17; 18-29; 30-45; 46-60; over 60
Condition road user	normal condition; drunken; sedated; ill

Factors road user	ignored red light; no priority, over white line; incorrect passing; sidestep maneuver; incorrect position on roadway; loss control steering wheel; no distance; fall
Factors vehicle	incorrect vehicle lights; bad tires; flat tire; defect trailer or cargo
Type of collision	multiple collision; frontal collision; parallel collision; sideways collision; pedestrian collision; collision obstacle on roadway; collision obstacle outside roadway; collision no obstacle
Type of obstacle	animal; train; streetcar; load on roadway; container; road works; street border; speed ramp; excavation; tree; public lighting; post; over crash barrier; against crash barrier; wall; fence; canal; other obstacle
Position pedestrian on footpath	pedestrian on cycle track; pedestrian out of vehicle; pedestrian right side roadway; pedestrian left side roadway; zebra crossing with traffic lights; zebra crossing with police officer; zebra crossing; next to zebra crossing with traffic lights; next to zebra crossing with police officer next to zebra crossing; no zebra crossing; pedestrian not moving on roadway
Visibility pedestrian	visible; not visible
Walking distance pedestrian	1-4 m; 5-10m; 11-15 m; over 16 m
Position cyclist	separated cycle track; marked cycle track on roadway; other cycle track
Cycle track	one way cycle track; two way cycle track normal direction; one way cycle track opposite direction

Gender passenger	male passenger; female passenger
Injuries passenger	dead; seriously injured; lightly injured
Position passenger	front seat; Back seat
Age passenger	0-17; 18-29; 30-45; 46-60; over 60 years old
Gender victim	male victim; female victim
Age victim	0-17; 18-29; 30-45; 46-60; over 60 years old
Injuries victim	dead; seriously injured; lightly injured
Number of road users	0; 1; 2; 3; 4; 5; 6; 7; 8
Number of passengers	0; 1; 2; 3; 4; 5
Number of victims	0; 1; 2; 3; 4; 5
Total number of lightly injured	0; 1; 2; 3; 4; 5; 6; 7
Total number of seriously injured	0; 1; 2; 3; 4; 5
Total number of deaths	0; 1; 2; 3; 4; 5

Samenvatting

In Vlaanderen worden momenteel 1014 ongevallocaties als ‘gevaarlijk’ beschouwd. Deze gevaarlijke plaatsen of zogenoemde ‘zwarte punten’ worden geselecteerd op basis van hun historische ongevallendata. Meer bepaald wordt voor elke locatie waar in de voorbije 3 jaar, 3 of meer letselongevallen plaats vonden een combinatie van gewichten gebruikt om de gevaarlijke ongevallenlocaties te rangschikken en te selecteren: respectievelijk 1 voor elke licht gewonde, 3 voor elke zwaar gewonde en 5 voor elke dode (combinatie 1_ 3_ 5). Indien de resulterende score minstens 15 bedraagt, wordt een ongevallenlocatie als een gevaarlijk punt beschouwd.

In dit doctoraat wordt aan de hand van een sensitiviteitsanalyse onderzocht wat de implicaties zijn van het gebruik van deze ranking methode op de selectie van de meest gevaarlijke ongevallocaties. Daarnaast worden, met behulp van data mining en statistische technieken, ongevallocaties geprofileerd in termen van ongevaldata en locatiekarakteristieken. Dit laat toe ongevalpatronen te identificeren en vervolgends gevaarlijke ongevallocaties aan te pakken met gerichte maatregelen. Daarnaast worden een aantal nieuwe modellen ontwikkeld die meer inzicht geven in de verschillende factoren die het al dan niet gebeuren van verkeersongevallen trachten te verklaren. Tot slot wordt in dit onderzoek nagegaan of gevaarlijke ongevallocaties in Vlaanderen zich verplaatsen in de tijd.

In hoofdstuk 3 leggen we uit dat verschillende prioriteiten ten aanzien van het verkeersveiligheidsbeleid kunnen vertaald worden in verschillende wegingsfactoren. De resultaten van een sensitiviteitsanalyse tonen aan dat er wel degelijk belangrijke gevolgen zijn voor het selecteren en rangschikken van gevaarlijke punten wanneer andere combinaties van gewichten worden gehanteerd. Bovendien heeft een verandering in de wegingsfactoren niet alleen een effect op het aantal ongevallocaties dat wijzigt bij de selectie van de meest gevaarlijke punten, het heeft ook een invloed op het type van locatie dat geselecteerd wordt en bijgevolg ook op toekomstige acties om deze locaties verkeersveilig te maken. Naast de impact van de gewichten wordt onderzocht welke impact het aantal passagiers in een voertuig heeft op de rangschikking van ongevallocaties. Deze resultaten tonen aan dat het toekennen van gewichten aan de ernst van het ongeval in plaats van aan alle gewonde inzittenden een belangrijk effect heeft op de selectie en rangschikking van gevaarlijke ongevallocaties. De overheid moet dan ook zorgvuldig beslissen of ze locaties willen rangschikken aan de hand van de ernst van het ongeval of de ernst en het aantal van alle gewonde inzittenden. Bovendien wordt in dit hoofdstuk nagegaan wat het effect is op de rangschikking van de ongevallocaties wanneer gewerkt wordt met het verwacht aantal ongevallen, geschat op basis van een hiërarchisch Baysiaans model, in plaats van met de historische ongevallendata. Immers, wanneer locaties geselecteerd worden als zijnde gevaarlijk op basis van hun aantal geobserveerde ongevallen dan kan 'bias by selection' verwacht worden. Dit houdt in dat bepaalde locaties, die op basis van hun historisch aantal ongevallen schijnbaar gevaarlijk zijn, misschien onterecht geselecteerd worden voor behandeling terwijl echte gevaarlijke locaties over het hoofd worden gezien. Verder ontwikkelden we een model om Baysiaanse ranking plots te genereren die gebruikt kunnen worden door het beleid om aan de hand van een grafisch instrument gevaarlijke ongevallocaties te selecteren op basis van een statistisch onderbouwde methode. Tot slot, bekijken we in een case study het gezamelijk effect van bovenstaande aspecten door enkel de ernstigste graad van verwonding per ongeval in rekening te brengen en gebruiken we gewichten voor deze type verwondingen die gebaseerd zijn op directe kosten, indirecte kosten en de waardering van een mensenleven. Dit resulteert respectievelijk in de gewichten 1_ 7_ 33 wanneer het ergste slachtoffer een licht, zwaar of dodelijk gewonde betreft. Daarnaast maken we gebruik van ranking plots, om de kans dat een locatie als gevaarlijk wordt gerangschikt visueel voor te stellen. Resultaten tonen aan dat de combinatie van deze 3 alternatieve rangschikkingcriteria een groot effect heeft op de selectie en rangschikking van gevaarlijke ongevallocaties. Op basis van deze resultaten willen we het beleid dan ook sensibiliseren om de criteria om ongevallocaties te rangschikken en selecteren zorgvuldig uit te kiezen. In deze context ontwikkelden we ook een mixed integer programming model om de optimale gewichten te genereren voor de 800 ongevallocaties die in de eerste plaats verantwoordelijk zijn voor het menselijk leed op de wegen. De resultaten van dit onderzoek tonen aan dat om de ongevallocaties te selecteren die het menselijk leed op de wegen maximaliseren het niet enkel voldoende is om de verschillende letselstypes te wegen, hetzij uit een ethish of economisch oogpunt, maar dat ook de relatie tussen het aantal verwondingen van de verschillende letseltypes moet in rekening genomen worden.

In hoofdstuk 4 maken we gebruik van een associatie algoritme op een data set van verkeersongevallen om gevaarlijke ongevallocaties te profileren in termen van ongevaldata en locatiekarakteristieken. We tonen aan dat frequent item sets en associatieregels gegenereerd kunnen worden om ongevalomstandigheden te identficeren die vaak samen voorkomen. Meer bepaald wordt deze techniek hier gebruikt om vaak voorkomende ongevaltypes te onderscheiden, elk met verschillende relevante ongevalkarakteristieken. Daarnaast identificeren we aan de hand van deze techniek ongevalpatronen die vaak voorkomen op locaties waar veel ongevallen gebeuren en vergelijken deze met locaties waar weinig ongevallen gebeuren. Bovendien tonen we aan dat ruimtelijke concentraties van verkeersongevallen gekarakteriseerd worden door specifieke omstandigheden, welke bijgevolg ook specifieke maatregelen vereisen om het aantal ongevallen op deze locaties te doen verminderen.

Tot slot gebruiken we in hoofdstuk 5 van deze thesis de Vlaamse ongevaldata van 1997-1999 en 1999-2001 om te onderzoeken of de gevaarlijke punten zijn verschoven in de tijd. Resultaten tonen aan dat, op basis van de huidige gebruikte Vlaamse definitie van een gevaarlijk punt, bijna de helft van de gevaarlijke punten uit de periode 1997-1999 nog steeds als gevaarlijk moet beschouwd worden in 1999-2001. Bovendien tonen we aan de hand probabilistische rankings aan dat bij de interpretatie van deze resultaten rekening moet gehouden worden met het regression to the mean effect en verschuiving van gevaarlijke locaties naar nabijgelegen locaties.