ELSEVIER

Accident Analysis and Prevention 37 (2005) 787–799

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Understanding spatial concentrations of road accidents using frequent item sets

Karolien Geurts [a], Isabelle Thomas [b], Geert Wets [a,*]

[a] *Transportation Research Institute Limburgs Universitair Centrum, Universitaire Campus, gebouw D, B-3590 Diepenbeek, Belgium*
[b] *National Fund for Scientific Research and Université catholique de Louvain, Department of Geography, Place Louis Pasteur 3, B-1348 Louvain-la-Neuve, Belgium*

Received 4 January 2005; accepted 23 March 2005

Abstract

This paper aims at understanding why road accidents tend to cluster in specific road segments. More particularly, it aims at analyzing which are the characteristics of the accidents occurring in "black" zones compared to those scattered all over the road. A technique of frequent item sets (data mining) is applied for automatically identifying accident circumstances that frequently occur together, for accidents located in and outside "black" zones. A Belgian periurban region is used as case study. Results show that accidents occurring in "black" zones are characterized by left-turns at signalized intersections, collisions with pedestrians, loss control of the vehicle (run-off-roadway) and rainy weather conditions. Accidents occurring outside "black" zones (scattered in space) are characterized by left turns on intersections with traffic signs, head-on collisions and drunken road user(s). Furthermore, parallel collisions and accidents on highways or roads with separated lanes, occurring at night or during the weekend are frequently occurring accident patterns for all accident locations. These exploratory results show the potentiality of the frequent item set method in addition to more classical statistical techniques, but also suggest that there is no unique countermeasure for reducing the number of accidents.

Keywords: Data Mining; Frequent Sets; "black" zones; Accidents; Belgium; Periurban

1. Introduction

Traffic collisions remain one of the leading causes of premature death and morbidity in most countries. In Belgium as in many European countries, traffic safety is currently one of the government's priorities. Identifying dangerous accident locations and profiling them in terms of accident-related data and location/environmental characteristics provide new insights into the complexity and causes of road accidents.

Long ago, the spatial structure of road accidents was demonstrated, but no official and universal agreement exists for defining significant spatial concentrations of road accidents. In general, methods developed for identifying accident 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 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 paper, the location and the length of the "black" zones are defined by means of local spatial autocorrelation

* Corresponding author. Tel.: +32 11 26 87 57/86 08; fax: +32 11 26 87 00.
E-mail address: geertwets@luc.ac.be (G. Wets).

0001-4575/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2005.03.023

indices (see Section 3.2), and they are considered as given in our problem. Therefore, the problem tackled in this paper is not the definition of the "black" zone, but its exploration. We argue that, indeed, it is not possible to develop effective countermeasures to reduce the number of accidents at these locations without being able to properly and systematically relate accident frequency and severity to a number of variables such as roadway geometries, traffic control devices, roadside features, roadway conditions, driver behavior or vehicle type (Kononov and Janson, 2002). Hence, several attempts are found in literature for explaining the spatial variation of road unsafety at several levels of spatial aggregation (see Flahaut, 2004a,b for a review). Our approach, however, is purely exploratory, 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. This data mining technique automatically identifies accident circumstances that frequently occur together. This way, we expose a number of hypotheses, which we then try to explain using other research studies and domain knowledge. Statistical models have been widely used on such 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). 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 datasets 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 large amounts of 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). However, in contrast with statistical techniques, 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 with 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 (Mannila, 2000). Additionally, in contrast with statistics, data mining is typically a form of secondary data analysis: the data has been collected for some other purpose than for answering a specific data analytical question. For the purposes of this paper it is sufficient to point out that statistical models are particularly likely to be preferable when fairly simple models are adequate and the important variables can be identified before modeling. However, when dealing with a large and complex data set of road accidents, the use of data mining methods seems particularly useful.

In literature some examples of the use of data mining in road accidents analyses 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 rule sets representing interesting subgroups in accident data (see e.g. Kavsek et al., 2002). Furthermore, decision trees (see e.g. Strnad et al., 1998; Clarke 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. Zeitouni and Chelghoum, 2001) can be applied.

In this research, data mining is applied for understanding the characteristics of the accidents associated to "black" zones or hazardous road segments. In particular, an existing technique of frequent item sets is used as an explorative technique to generate accident patterns, which can give rise to possible new and surprising accident patterns that were not yet found in other research. More specifically, 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.

The remainder of this paper is organized as follows. First a formal introduction to the association algorithm and the concept of frequent item sets is provided (Section 2). This will be followed by a description of the dataset and the studied area (Section 3). In Section 4, the empirical study is explained and in Section 5 the results of this study are presented. The paper will be completed with a summary of the conclusions and directions for future research.

2. Frequent Item sets

2.1. KDD process

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

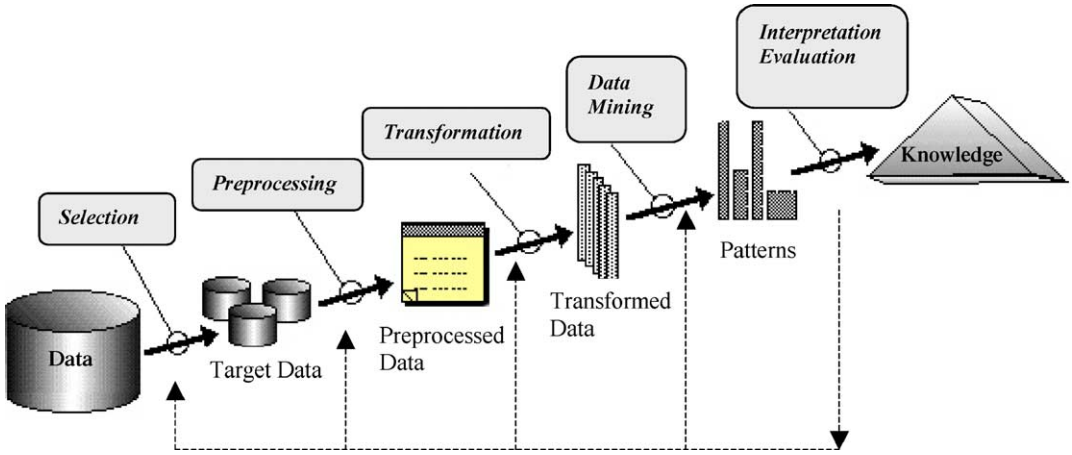

Fig. 1. The KDD-process (Fayyad et al., 1996).

(KDD) process (see Fig. 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. A more detailed description of these different data preparation steps for data mining can be found in Pyle (1999).

Depending on the objectives of the research, different data mining techniques can be implemented. In general, two major categories of data mining tasks can be discerned (Berry and Linoff, 1997): prediction and description. For example, classification and estimation tasks can be seen as prediction tasks: the user wants to predict the (discrete or continuous) value of an unknown attribute. 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, the technique of frequent item sets is used to perform the data mining task of affinity grouping. This will be explained in the following section (see Section 2.2). However, the choice for one particular data mining technique 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 in order to choose the appropriate technique. For example, the association algorithm is able to identify all the accident circumstances that frequently occur together. However, this algorithm cannot give any explanation about the causality of these accident patterns. Furthermore, to evaluate and interpret the interestingness of the results domain knowledge or the use of additional statistical techniques is essential.

2.2. Association algorithm

In this study, an association algorithm is used to obtain a descriptive analysis of “black” zones. 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 its 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). More specifically, the association algorithm produces frequent item sets describing underlying patterns in data. In contrast to predictive accident models, the strength of this algorithm lies within the identification of accident circumstances that frequently occur together (Geurts et al., 2003)). 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 that the user-specified minimum support (minsup). These items are considered to be “frequent” item sets.

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

Table 1
Interpretation of Lift

Outcome	Interpretation
$+\infty > L > 1$	Positive interdependence effects between A and B
$L = 1$	Conditional independence between A and B
$0 < L < 1$	Negative interdependence effects between A and B

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 $\{C\}$ are both 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.

2.3. Interesting patterns

The association algorithm generates all item sets that have support higher than the minimum support value. However, earlier research showed that a large number of the generated item sets will be trivial and thus a filter is needed to post-process the discovered item sets to retain the interesting ones. Two properties of the association algorithm can be used to distinguish trivial from non-trivial patterns. A first, more formal method (Brin et al., 1997)) to assess the dependence between the items in the item set is lift (L).

Definition 1. Lift (L)

$$L = \frac{s(A,B)}{s(A)\cdot s(B)}$$

The nominator $s(A, B)$ measures the observed frequency of the co-occurrence of the items A and B. The denominator $s(A)\cdot(B)$ 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 1 illustrates the three possible outcomes for the lift value and their associated interpretation for the dependence between the items.

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; Geurts et al., 2003).

Definition 2. Interest (Int)

$$\text{Int} = \frac{s_b - s_n}{\max\{s_b, s_n\}}$$

This interestingness measure is based on the deviation in support values of the frequent item sets discovered for the accidents that occurred within a "black" zone from the accidents that occurred outside a "black" zone. The nominator $s_b - s_n$ measures the difference in support for the accident characteristics in the "black" zones (s_b) and non-"black" zones (s_n). The expression $\max\{s_b, s_n\}$ is called the normalizing factor as it normalizes the interestingness measure onto the scale $[-1, 1]$. Since in this research, we are mainly interested in profiling the "black" zones, we will pay special attention to the item sets with a positive interest value, i.e. approximating "1".

2.4. Example

For example, consider the following accident data containing three accidents:

Accident: Accident circumstances
Accident 1: Rain, crossroad, traffic lights
Accident 2: Rain, crossroad, traffic signs
Accident 3: Normal weather, zebra crossing, pedestrian

Suppose we set the minimum support value (minsup) = 60%. This means that the accident characteristics should occur in at least 60% of all the accidents in order to be considered as frequent. This leads to the following results:

- Frequent item sets of size 1: s (rain) = 2/3 (66.6%) s (crossroad) = 2/3 (66.6%); Frequent item sets of size 2: s (rain, crossroad) = 2/3 (66.6%).
- Lift (rain, crossroad) = s (rain, crossroad)/(s (rain)·s (crossroad)) = (2/3)/[(2/3)·(2/3)] = 3/2 (>1) ⇒ positive interdependence between rain and crossroad.

  Suppose the item set (rain, crossroad) is also frequent in a second data set with: (rain, crossroad) = 3/4 (75%).
- Interest = s_1 (rain, crossroad) − s_2 (rain, crossroad)/max $\{s_1, s_2\}$ = [(2/3) − (3/4)]/(3/4) = −1/9 (<1) ⇒ this value is close to "0" indicating that although this item set is very descriptive for both data sets (lift > 1), this item set is not very discriminating between the two data sets.

3. The data set

3.1. The studied area

In Belgium, each road accident occurring on a public road and involving casualties is reported officially (National Institute of Statistics). Its location is known accurately on numbered roads because there is a stone marker at every hectometer; numbered roads are motorways, national and provincial roads linking towns together. Hence, this analysis is limited to accidents with casualties on numbered roads. 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.

Belgium is a quite small country but densely built and mainly urbanized (10.3 millions inhabitants; 30,528 km^2). However, large disparities exist within the country (see e.g. Merenne et al., 1997). Hence, our analysis here is limited to one administrative region: the Walloon Brabant, which

is a province extending South of Brussels. It has a surface area of 1100 km^2 and counts almost 350,000 inhabitants and 4604 km 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 (Thomas et al., submitted for publication). The Eastern part is still rural, the Western part more industrial. Limiting the extent of the studied area enables one to better control other sources of variations (mobility habits, friction of distance, mobility policies, etc).

Traffic accident data are obtained from the "Belgian Analysis Form for Traffic Accidents" that needs to be filled in by a police officer for each road accident that occurs on a public road and that involves casualties. 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, . . .).

These data indicate that for the region of Walloon Brabant, 1861 injury accidents occurred between 1997 and 1999. In these accidents, 81 persons were deadly injured, 333 seriously and 2374 lightly injured.

Furthermore, an initial analysis on these data indicated that 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. For example, 73% of the accidents in the dataset occurred under normal weather conditions. As explained earlier, this will have no effect however on the validity of the results since the association algorithm produces the lift value that corrects the importance of each rule by taking the frequency of the attributes in the dataset into account.

3.2. Definition of the "black" zones

The location and the extension of hazardous locations are defined in a preceding study (Flahaut et al., 2003). Simply said, the hectometer (100 m) 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 then gives 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). Here we used a local index developed as a local indicator of spatial autocorrelation (LISA) by Anselin, 1995. It takes high values for hectometers located close to each other and having large numbers of accidents. Closeness is measured here 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 et al., 2004). These sensitivity analyses confirm the performance of the technique as well as a strong spatial structure of the road accidents in the Brabant Walloon area. This means that the location of the road accident 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 paper. We refer to former studies for a critical analysis of the method (Flahaut, 2004a,b; Flahaut et al., 2003; Flahaut and Thomas, 2002).

From a former analysis on the same studied area (Eckhardt et al., 2004) we know that 47% of the road hectometers never registered 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 km 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 dataset, containing the accidents that took place outside a "black" zone involves 1287 road accidents (note: the belonging or not of an accident location to a "black" zone could not be defined for 21 accidents).

4. Empirical study

As explained in Section 2.1 of this paper, we can distinguish different steps in the mining process: a pre-processing step and a transformation step in which the available data are prepared for the use of the mining technique, a mining step for generating the frequent item sets and a post-processing

step for evaluating and interpreting the most interesting patterns.

4.1. Pre-processing and 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 (see Section 3). Some variables have a continuous character. Discretization of these variables is necessary, since generating frequent item sets requires a data set for which all items are discrete. The intervals for these variables were created on the basis of expert knowledge on traffic safety issues such as traffic rush hours, or types of road users (drivers license) in Belgium. For example, six new attributes were created from the continuous variable "time of accident": morning rush hour (7–9 a.m.), morning (10–12 a.m.), afternoon (1–3 p.m.), evening rush hour (4–6 p.m.), evening (7–9 p.m.) and night (10–6 a.m.). A second example includes the variable age for which six new intervals were created: age between 0 and 17, 18 and 29, 30 and 45, 46 and 60 and age over 60. For variables where no domain knowledge for grouping the attribute values could be found, we used the equal frequency binning method; this is a discretization method that generates intervals containing an equal number of observations (Holte, 1993). Next, some new items will be created based on existing information in the data set. For example, the variable week/weekend is created from the variables time and date using the expert knowledge that the weekend starts on Friday evening at 7 p.m. and ends on Monday morning at 6 a.m.

Furthermore, attributes with nominal values have been transformed into binary attribute values. This means that dummy variables were created by associating a (0,1) attribute to each nominal attribute value. Finally, irregularities such as data inconsistencies, missing values, redundant variables and double counts are tracked, listed and removed from the data sets (Casaer et al., in preparation).

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 A). 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.

4.2. Mining for frequent items

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. 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.

4.3. Post-processing the frequent item sets

As stated in the introduction of this paper, 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 (see Section 2). 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 Definition 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.

Table 2
Frequent item sets for accidents in 'black' zones

N	Item 1	Item 2	Item 3	Item 4	Support (%)	Lift
1	Near intersection	Signalized intersection			9.4	4.09
2	Signalized intersection	Left turn			5.4	3.54
3	Near intersection	Signalized intersection	Left turn		5.4	14.51
4	District or province road	Signalized intersection	Left turn		5.2	5.87
5	Straight direction	Signalized intersection	Left turn		5.1	5.17
6	Two road users	No priority	Sideways collision	Left turn	5.9	15.36
7	Rain	Aquaplaning			5.1	3.42
8	Wet road surface	Rain	Aquaplaning		5.1	7.88
9	Built-up area	Pedestrian collision			5.6	2.74
10	Road with one road lane	Speed 50 km h	Built-up area	Pedestrian collision	5.6	16.80
11	Built-up area	Age road user 0–17			5.4	2.73
12	One road user	Loss control to the left	Against crash barrier		5.2	7.96
13	One road user	Collision obstacle outside roadway	Loss control to the left	Against crash barrier	5.2	26.37
14	Speed 120 km h	Loss control to the left	Against crash barrier		5.1	6.53
15	Speed 120 km h	Collision obstacle outside roadway	Loss control to the left	Against crash barrier	5.1	21.65
16	Road user normal condition	Car	Speed 50 km h	Loss control steering wheel	5.2	0.49

5. Results

5.1. Accident patterns in "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 2 gives an overview of the most interesting of these frequent item sets. In the remainder of this paper, 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 2]. More specifically, accidents on intersections covered by traffic lights frequently coincide with a road user making a left turn [2,3,4,5]. 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 is really characteristic of black zones and absolutely not of 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. These results can be linked with the results obtained by Larsen and Kline (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). Even if obvious, there is however no simple cause and effect relationship. It has already been widely demonstrated that during wet conditions accident numbers increase (see e.g. Brodsky and Hakker, 1988). This is partly due to 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". Table 2 mentions rain as well as aquaplaning 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 three 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 (see Table 3) 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 (infrastructural and/or behavioral).

Next, Table 2 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

Table 3
Frequent item sets for accidents outside 'black' zones

N	Item 1	Item 2	Item 3	Item 4	Support (%)	Lift
17	Intersection traffic signs	Left turn			7.4	2.09
18	Near intersection	Left turn	Intersection traffic signs		7.4	12.42
19	Two road users	Near intersection	Left turn	Intersection traffic signs	6.5	20.48
20	Speed 50 km h	No priority	Intersection traffic signs		5.6	6.45
21	Speed 50 km h	Near intersection	No priority	Intersection traffic signs	5.6	27.59
22	Speed 50 km h	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

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 et al., 1991; Braddock et al., 1994; LaScala et al., 2000 or Julien and Carré, 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 Walloon (periurban) 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 crossings with traffic lights are rare.

Furthermore, for some hazardous road segments, road users frequently lose control of 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/h [14,15] and less frequently on a road with a speed limit of 50 km/h [16]. Given the 120 km/h limit, these accidents are located on highways. Run-off-roadway accidents have already been studied often (see e.g. literature review in Lee and Mannering, 2002). They are often related to an inadequacy of 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 paper. 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.

5.2. 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 eight 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 3 gives an overview of the most interesting of these frequent item sets.

In contrast to the results for accidents in "black" zones, Table 3 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 km/h where no priority is given, resulting in a side impact collision [20–22]. In combination with the results of Table 2 these patterns indicate that a strong difference exists 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 modern traffic lights or roundabouts.

Next, Table 3 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 two 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.

Table 4
Frequent item sets for all accidents (belonging or not to a 'black' zone)

N	Item 1	Item 2	Item 3	Item 4	S_b (%)	Lift_b	S_n (%)	Lift_n	Int
26	Night	Highway			12.3	1.39	5.6	1.391	0.54
27	Public lighting	Night	Road with separated road lanes		13.4	3.78	6.2	3.777	0.54
28	Highway	Weekend			16.8	1.15	7.4	1.152	0.56
29	Parallel collision	Highway			13.4	1.18	6.3	1.185	0.53
30	Parallel collision	Road with separated road lanes	Highway		13.4	2.09	6.1	2.087	0.54
31	Rain	Speed 120 km h			10.8	1.09	5.0	1.098	0.53
32	Wet road surface	Road with separated road lanes	Rain		17.9	2.66	5.9	2.66	0.67
33	Wet road surface	Outside built-up area	Road with separated road lanes	Rain	16.9	3.73	5.7	3.735	0.66
34	Car	Wet road surface	Road with separated road lanes	Rain	5.1	3.04	17.2	3.036	0.70
35	Car	Public lighting	Road with separated road lanes	Rain	14.2	2.34	5.2	2.345	0.63

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 needed here to further understand the head-on accidents. Additionally, with one road user driving in a positive way and another road user driving in a negative way, the accident will take place on a wet road surface less frequently than expected [24]. These results can probably be explained by the fact that aquaplaning does not frequently occur outside "black" zones (see Section 5.1). 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 3 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.

5.3. Common accident patterns

In Sections 5.1 and 5.2, 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 strong in both data sets. Therefore, we can use the interest value Int (see Definition 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 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 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% ($S_b = 12.3\%$) [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% ($S_n = 5.6\%$) [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 km/h [30] and outside the built-up area [31]. These results can be related to the figures in Table 2, 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 in

Section 5.1, wet conditions are less manageable by road users in "black" zones, for one reason or another (infrastructural or behavioral).

6. Discussion

6.1. Frequent item sets and accident analysis

In this paper, the association algorithm was used on a data set of road accidents to profile "black" zones in terms of accident-related data and location characteristics. More specifically, frequent item sets are generated to identify accident circumstances that frequently occur together in order to find out which factors explain the occurrence of the accidents in "black" zones. As explained in the introduction, the use of this technique coincides with the explorative character of this research since it describes the co-occurrence of accident circumstances and gives direction to more profound research on the causes of these accident patterns and explanation. 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 descriptive analysis of accident patterns inside "black" zones, it also creates the possibility to find the accident characteristics that are frequent for all accidents but that occur more frequently inside than outside "black" zones.

6.2. Accidents patterns in "black" zones

The most important result of this research is that road accident 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. Based on the results of an in-depth analysis of left-turn collisions by Larsen and Kline (2002), we could conclude that better signalized intersections could be a short-term solution for these types of accidents. In the long term, one should consider the use of roundabouts to increase traffic safety in these "black" zones.

A second important accident circumstance both inside and outside a "black" zone is rainy weather conditions. This is partly due to slippery roads (Brodsky and Hakker, 1988) but also to risk-taking behavior of drivers (Edwards, 1996). 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. Hence education and environmental prevention efforts should 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 inappropriate 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).

The findings of this paper are rather suggestive and limited to one data set. Furthermore, in this paper the location and the length of the "black" zones are defined by means of local spatial autocorrelation indices (see Section 3.2) and are considered as given in our problem. Results of this paper 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.

6.3. Future studies

Although the analysis carried out in this paper reveals several interesting accident patterns, which, in turn, provide valuable input for purposive government traffic safety actions, several issues remain for future research. First, the skewed character of the accident data limits the amount of information contained in the dataset and will therefore restrict the number of circumstances that will appear in the results. Moreover, the variables used here are restricted to those collected by the police. We intend to extend our range of variables to other sources in the future (traffic, land-use, accessibility, etc.) in order to better approach the explanation process. Secondly, 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 frequent item sets to discover the most interesting accident patterns. Furthermore, using other

discretization categories (e.g. for the variable age) and creating more new variables (e.g. type of accident: car-car, car-cyclist, . . .) instead of using the existing variables could lead to other associations. Note, however, that omitting items and at the same time not creating any new items will not cause the algorithm to reveal new accident patterns, since the association algorithm will in any case generate all patterns that are frequent in the data set. Finally, the identified interesting accident patterns can be used in statistical models to test their significance and to evaluate the difference in importance of these effects on the occurrence of "black" zones.

Acknowledgements

This research was supported by the OSTC and the Flemish Research Centre for Traffic Safety. The authors would also like to thank dr. Tom Brijs for his encouragement and helpful suggestions.

Appendix A

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; intersection; signalized intersection flashing light; intersection traffic signs; intersection priority to right
Location characteristic	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 centre; 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 a.m.); morning (10–12 a.m.); afternoon (1–3 p.m.); evening rush hour (4–6 p.m.); evening (7–9 p.m.); night (10–6 a.m.)
Type of road user	Car; car double use; minibus; light truck; camper; truck; truck and trailer; truck; tractor; bus; trolleybus; motor coach; motorbike under 400 cm^3; motorbike over 400 cm^3; moped A; moped B; moped three to four 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
Sex road user	Male road user; female road user
Consequences road user	Dead road user; seriously injured road user; lightly injured road user; uninjured road user
Age road user	Age road user 0–17; 18–29; 30–45; 46–60; over 60
Condition road user	Road user in normal condition; drunken road user; sedated road user; ill road user
Factors road user	Through 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	Pedestrian visible; pedestrian not visible
Walking distance pedestrian	Walking distance 1–4 m; walking distance 5–10 m; walking distance 11–15 m; walking distance 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
Consequences passenger	Dead passenger; seriously injured passenger; lightly injured passenger

Appendix A (*Continued*)

Variable	Item
Position passenger	Passenger front seat; passenger 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.
Consequences victim	Dead; seriously injured; lightly injured
Number of road users	0; 1; 2; 3; 4; 5; 6; 7; 8 road users
Number of passengers	0; 1; 2; 3; 4; 5 passengers
Number of victims	0; 1; 2; 3; 4; 5 victims
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

References

Agent, K.R., Deen, R.C., 1975. Relationship between roadway geometrics and accidents. Transportation Research Record 541, Washington, DC.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data, Washington, DC, USA, May 26–28, pp. 207–216.

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, pp. 307–328.

Anand, S.S., Bell, D.A., Hughes, J.G., Patrick, A., 1997. Tackling the cross sales problem using data mining. In: Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining.

Anselin, L., 1995. Local indicators of spatial association-LISA. Geographical Anal. 27 (2), 93–115.

Baker, S., Waller, A., Langlois, J., 1991. Motor vehicle deaths in children: geographic variations. Accid. Anal. Prev. 23, 19–28.

Berry, M., Linoff, G., 1997. Data Mining Techniques for Marketing, Sales and Customer Support. John Wiley & Sons.

Braddock, M., Lapidus, G., Cromley, E., Cromley, R., Burke, G., Banco, L., 1994. Using a geographic information system to understand child pedestrian injury. Am. J. Public Health 84, 1158–1161.

Brin, S., Motwani, R., Silverstein, C., 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, pp. 265–276.

Brodsky, H., Hakker, A.S., 1988. Risk of a road accident in rainy weather. Accid. Anal. Prev. 20, 161–176.

Casaer, F., Eckhardt, N., Steenberghen T., Thomas, I., Wets, G., Quality assessment of the Belgian traffic accident data, Paper in progress.

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

Clarke, R., Forsyth, Wright, R., 1998. Machine learning in road accident research: decision trees describing road-accidents during cross-flow turns. Ergonomics 41 (7), 1060–1079.

Dekeersmaecker, M.-L., Frankhauser, Thomas, I., 2004. Analyse de la réalité fractale périurbaine: l'exemple de Bruxelles. L'Espace Géographique 3, 219–240.

Doherty, S.T., Andrey, J.C., MacGregor, C., 1998. The situational risks of young drivers: the influence of passengers, time of day, and day of week on accident rates. Accid. Anal. Prev. 30 (1), 45–52.

Eckhardt, N., Flahaut, B., Thomas, I., 2004. Spatio-temporalité des accidents de la route en périphérie urbaine L'exemple de bruxelles. Recherche Transports et Sécurité 82, 35–46.

Edwards, J., 1996. Weather-related road accidents in England and Wales: a spatial analysis. Accid. Anal. Prev. 4 (3), 201–212.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, pp. 1–34.

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying 'black' zones. A comparative approach. Accid. Anal. Prev. 35 (6), 991–1004.

Flahaut, B., Thomas, I., 2002. Identifier les zones noires d'un réseau routier par l'autocorrélation spatiale locale. Revue Internationale de Géomatique 12 (2), 245–261.

Flahaut, B., 2004a. Impact of infrastructure and local environment on road insecurity. Logistic modeling with spatial autocorrelation. Accid. Anal. Prev. 36, 1055–1066.

Flahaut, B., February 2004b. Towards a sustainable road safety in Belgium. Location of spatial concentrations of road accidents and explanatory modelling. Ph.D. dissertation. Department of Geography, Louvain-la-Neuve, 107 pp.

Friedman, J.H., 1997. Data mining and statistics: what's the connection? In: Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics.

Geurts, K., Wets, G., Brijs, T., Vanhoof, K., 2003. Profiling high frequency accident locations using association rules. In: Proceedings Transportation Research Board (CD-ROM), Washington, DC, USA, January 12–16.

Greibe, P., 2003. Accident prediction models for urban roads. Accid. Anal. Prev. 35, 273–285.

Hauer, E., 1996. Identification of sites with promise. In: Transportation Research Record 1542. 75th Annual Meeting, Washington, DC, pp. 54–60.

Hipp, J., Güntzer, U., Nakhaeizadeh, G., 2000. Algorithms for association rule mining—a general survey and comparison. SIGKDD Explorations 2 (1), 58.

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. Machine Learn. 11, 63–90.

Hosking, J., Pednault, E., Sudan, M., 1997. A statistical perspective on data mining. Future Gen. Comput. Syst. 13, 117–134.

Joly, M.-F., Bourbeau, R., Bergeron, J., Messier, S., 1992. Analytical approach to the identification of hazardous road locations: a review of the literature. Centre de recherche sur les transports, Université de Montréal.

Julien, A., Carré, J.-R., 2002. Cheminements piétonniers et exposition au risque. Recherche Transports Sécurité 76, 173–189.

Kavsek, B., Lavrac, N., Bullas, J.C., 2002. Rule induction for subgroup discovery: a case study in mining UK traffic accident data. In: Proceedings of Conference on Data Mining and Warehouses (SiKDD2002), Ljubljana, Slovenia, October 15.

Kononov, J., Janson, B., 2002. Diagnostic methodology for the detection of safety problems at intersections. In: Proceedings of the Trans-

portation Research Board (CD-ROM), Washington DC, USA, January 13–17.

Larsen, L., Kline, P., 2002. Multidisciplinary in-depth investigations of head-on and left-turn road collisions. Accid. Anal. Prev. 34, 367–380.

LaScala, E., Gerber, D., Gruenewald, P., 2000. Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accid. Anal. Prev. 32, 651–658.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34, 149–161.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. In: Proceedings of the Transportation Research Board (CD-ROM), Washington, DC, USA, January 13–17.

Levine, N., 2002. CrimeStat II : A Spatial Statistics Program for the Analysis of Crime Incident Locations (version 2.0). Ned Levine & Associates: Houston, TX/National Institue of Justice, Washington, DC.

Ljubic, P., Todorovski, L., Lavrac, N., Bullas, J.C., 2002. Time series analysis of UK traffic accident data. In: Proceedings of Conference on Data Mining and Warehouses (SiKDD 2002), Ljubljana, Slovenia, October 15.

Maher, M., 1990. A bivariate negative binomial model to explain traffic accident migration. Accid. Anal. Prev. 22 (5), 487–498.

Mannila, H., 2000. Theoretical frameworks for data mining. SIGKDD Explorations 1 (2), 30–32.

Martin, J.-L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prev. 34, 619–629.

Merenne, B., Van der Haegen, H., Van Hecke, E., 1997. La Belgique. Diversité territoriale. Bulletin du Crédit Communal, no. 202. Also available on Internet http://www.belspo.be.

Moran, P., 1948. The interpretation of statistical maps. J. R. Stat. Soc. 10b, 243–251.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accid. Anal. Prev. 31 (6), 705–718.

Nguyen, T.N., 1991. Identification of accident 'black' spot locations, an overview. VIC Roads/Safety Division, Research and Development Department, Australia.

Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufman, San Francisco, CA.

Rajalin, S., 1994. The connection between risky driving and involvement in fatal accidents. Accid. Anal. Prev. 26 (5), 555–562.

Silcock, D.T., Smyth, A.W., 1985. Methods of identifying accidents 'black' spots. Transport Operations Research Group, Department of Civil Engineering, University Of Newcastle Upon Tyne.

Strnad, M., Jovic, F., Vorko, A., Kovacic, L., Toth, D., 1998. Young children injury analysis by the classification entropy method. Accid. Anal. Prev. 30 (5), 689–695.

Taber, J.T., 1998. Multi-objective optimization of intersection and roadway design. Utah Transportation Center, Utah State University.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accid. Anal. Prev. 28, 251–264.

Thomas, I., Frankhauser, P., De Keersmaecker, M.-L. Fractal dimension versus density of the built-up surfaces in the periphery of Brussels. Refereed paper presented in Porto at the European Regional Science Association Congress, submitted for publication.

Vandersmissen, M.H., Pouliot, M., Morin, D.R., 1996. Comment estimer l'insécurité d'un site d'accident: état de la question. Recherche Transports Sécurité 51, 49–60.

Vistisen, D., 2002. Models and methods for hot spot safety work. Ph.D. dissertation. Denmark, 168 pp.

Wong, Y., Nicholson, A., 1992. Driver behaviour at horizontal curves: risk compensation and the margin of safety. Accid. Anal. Prev. 24 (4), 425–436.

Zeitouni, K., Chelghoum, N., 2001. Spatial decision tree – application to traffic risk analysis. In: ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, June 26–29.