The Use of Rule-Based Knowledge Discovery Techniques to Profile Black Spots

Karolien Geurts, Geert Wets, Tom Brijs and Koen Vanhoof
Limburg University Centre
Faculty of Applied Economics
University Campus – Diepenbeek
Belgium.

ABSTRACT

In Belgium, traffic safety is currently one of the highest topics on the list of priorities of the government. The identification of black spots and black zones and profiling them in terms of accident related data and location characteristics must provide new insights into the complexity and causes of road accidents which, in turn, provide valuable input for government actions. Data mining is the extraction of information from large amounts of data. The use of data mining algorithms is therefore particularly useful in the context of large datasets on road accidents. In this paper, association rules are used to identify accident circumstances that frequently occur together. The strength of this descriptive approach lies within the definition of different accident types and the identification of relevant variables that make a strong contribution towards a better understanding of accident circumstances. An analysis of the produced set of rules, describing underlying patterns in the data, indicates that five aspects of traffic accidents can be discerned: collision with a pedestrian, collision in parallel, sideways collision, week/weekend accidents and weather conditions. For each of these accident types, different variables play an important role in the occurrence of the accidents.

1 INTRODUCTION

In Belgium, every year approximately 50.000 injury accidents occur in traffic, with almost 70.000 victims, of which 1.500 deaths. In 1998 the probability of having a deadly accident was almost 35% higher in Belgium than the European average. Based on these figures, Belgium has a bad record towards traffic safety in comparison with most other European countries (Belgian Institute for Traffic Safety and National Institute for Statistics 2000). Not only does the steady increase in traffic intensity pose a heavy burden on the society in terms of the number of casualties, the insecurity on the roads will also have an important effect on the economic costs associated with traffic accidents. Accordingly, traffic safety is currently one of the highest topics on the list of priorities of the Belgian government.

Since a few decades, traffic accident data are registered and analysed to support the traffic safety policy. The identification of geographical locations with highly concentrated traffic accidents (black spots and black zones) and profiling them in terms of accident related data and location characteristics must therefore provide new insights into the complexity and criteria that play a significant role in the occurrence of traffic accidents to provide valuable input for government actions towards traffic safety. According to Kononov (2002) it is not possible to develop

effective counter-measures without being able to properly and systematically relate accident frequency and severity to roadway geometrics, traffic control devices, roadside features, roadway conditions, driver behaviour or vehicle type.

Lee, Saccomanno and Hellinga (2002) indicate that in the past, statistical models have been widely used to analyze road crashes and to explain the relationship between crash involvement and traffic, geometric and environmental factors. However, Chen and Jovanis (2002) demonstrate certain problems that may arise when using classic statistical analysis on datasets with large dimensions. This is where data mining comes into play. Data mining is 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 is therefore particularly useful in the context of large datasets on road accidents.

In this paper, the data mining technique of association rules is used to obtain a descriptive analysis of the accident data. In contrast with predictive models, the strength of this algorithm lies within the identification of relevant variables that make a strong contribution towards a better understanding of the circumstances in which the accidents have occurred which, in turn, facilitates the definition of different accident types. Hereby, the emphasis will not only lie on the acquired accuracy 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.

The paper is organized as follows. In section 2, a formal introduction to the technique of association rules is provided. Subsequently, section 3 presents a description of the dataset and the results of the empirical study. Finally, section 4 presents a summary of the conclusions and directions for future research.

2 ASSOCIATION RULES

Association rules is a data mining technique that can be used to efficiently search for interesting information in large amounts of data. More specifically, the association algorithm produces a set of rules describing underlying patterns in the data. Informally, the support of an association rule indicates how frequent that rule occurs in the data. The higher the support of the rule, the more prevalent the rule is. Confidence is a measure of the reliability of an association rule. The higher the confidence of the rule, the more confident we are that the rule really uncovers the underlying relationships in the data. It is obvious that we are especially interested in association rules that have a high support and a high confidence.

The concepts behind association rules and suggested algorithms for finding such rules were first introduced by Agrawal, Imielinski & Swami (1993). They provided the following formal description of this technique:

Let $I = \{i_1, i_2, \ldots, i_k\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its *TID*. We say that a transaction T *contains* X, a set of some items in I, if $X \subseteq T$. An *association rule* is an implication of

the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the transaction set D with *confidence* c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has *support* s in the transaction set D if s% of transactions in D contain $X \cup Y$. 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)*.

Generating association rules involves looking for so-called *frequent itemsets* in the data. Indeed, the support of the rule $X \Rightarrow Y$ equals the frequency of the itemset *{X, Y}*. Thus by looking for frequent itemsets, we can determine the support of each rule (Mannila 1997). The problem of discovering association rules can therefore be decomposed into two sub-problems:

1. Generating all itemsets that have a support higher than the user-defined minsup. These itemsets are called *frequent itemset*s.
2. Use this collection of frequent sets to generate the rules that have confidence higher than the user-defined minimum confidence.

DEFINITION 1 Frequency of an itemset
s(X, D) represents the frequency of itemset X in D, i.e. the fraction of transactions of D that contain X.

DEFINITION 2 Frequent itemset
An itemset X is called frequent in D, if s(X, D) $\geq \sigma$ *with* σ *the minsup.*

A typical approach (Agrawal et al. 1996) to discover all frequent sets X is to use the insight that all subsets of a frequent set must also be frequent. This insight simplifies the discovery of all frequent sets considerably, 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. This process is continued until no more candidate sets can be formed. Once all frequent sets are known, finding association rules is easy. Namely, for each frequent set X and each $Y \subset X$ verify whether the rule $X \setminus \{Y\} \Rightarrow Y$ 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 set.

3 EMPIRICAL STUDY

3.1 **Description of the Dataset**

This study is based on a large 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, the 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.

In total, 18.639 traffic accident records were available for analysis. However, in Belgium only the district and province roads are provided with kilometre and hectometre marks, what means that only the accidents that occurred at these roads can be located near such a mark. The accidents that take place at a non-numbered road can afterwards be located using the numbers of the houses in the street. Depending on the relation with a hectometre mark or not, the data set consists of two types of traffic accident data. Not only does the presence of a hectometre mark facilitate the locating of the accidents, it also gives additional information on the functional and physical characteristics of the road on which the accident has occurred. Since our main interest in this study lies within the identification and profiling of geographical locations, with as much relevant information as possible, where a high number of accidents occur, this analysis will concentrate on the accidents that can easily be located by the hectometre mark. Selecting these records from the data set resulted in a total of 10.672 traffic accident records.

Furthermore, the traffic accident data contain a rich source of information on the different 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 (fatigue, alcohol, …) and geographical conditions (location, functional and physical characteristics, …). In total, 84 attributes are included in the dataset.

An initial analysis on the dataset indicated that the 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, almost 99 % of the accidents in the dataset occurred in the inner city. As will be explained later (section 3.2.3), 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 dataset into account.

3.2 **Mining for Association Rules**

We distinguish three steps in the mining process: a preprocessing step in which the available data is prepared for the optimal use of the mining technique (section 3.2.1),

a mining step for generating the association rules (section 3.2.2) and a post-processing step for identifying the most interesting association rules (section 3.2.3).

3.2.1 *Preprocessing the Dataset*

To explore association relationships between traffic accident attributes, we did not use all the available traffic accident records. Since our prime interest lies in the identification and understanding of black spots and black zones, only the traffic accidents that occurred at a high frequency accident location were selected for the analysis. To identify these locations, a criterion of minimum ten accidents per location was used. 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).

Finally, 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.

3.2.2 *Generating Association Rules*

A minimum support of 5 percent 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 exponential growth of the number of items in the frequent itemsets. Accordingly, the number of rules that will be generated, will cause further research on these results to be impossible due to memory limitations. For example, for a minimum support of 1 percent the algorithm generated more tan 2 million association rules. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable of generating trivial rules.

From this analysis, with a minsup = 5 percent and minconf = 30 percent, the algorithm obtained 101.861 frequent itemsets of maximum size 4 for which 313.663 association rules could be generated. These rules are further processed to select the most interesting rules.

3.2.3 *Post-processing the Association Rules Set*

The association algorithm generates all rules that have confidence and support higher than minconf and minsup. This implies that all rules with high s(X) and high s(Y) will always be generated even if there is no statistically significant dependence between the *X* and *Y* itemsets. Therefore, a large subset of the generated rules set will be trivial. The purpose of post-processing the association rules set is to identify the subset of interesting (i.e., non-trivial) rules in a generated set of association rules.

Two properties of association rules can be used to distinguish trivial from non-trivial rules. A first, more formal method (Brin et al. 1997) to assess the dependence between the two itemsets in the association rule is *interest*.

DEFINITION 3 Interest

$$I = \frac{s(X \Rightarrow Y)}{s(X) * s(Y)}$$

The nominator s(X⇒Y) measures the observed frequency of the co-occurrence of the items in the antecedent (X) and the consequent (Y) of the rule. The denominator s (X) * s (Y) measures the expected frequency of the co-occurrence of the items in the antecedent and the consequent of the rule 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 interest measure and their associated interpretation for the dependence between the items in the antecedent and consequent of the rule.

Table 1: **Interpretation of interest**

Outcome	*Interpretation*
Interest> 1	Positive interdependence effects between *X* and *Y*
Interest = 1	Conditional independence between *X* and Y
Interest < 1	Negative independence effects between *X* and *Y*

A second method to define the interestingness of a rule is looking at the statistical rule significance (Silverstein, Brin and Motwani 1998).

DEFINITION 4 Statistical Rule Significance

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).

The statistical rule significance is determined using the χ^2- test for statistical independence and can be negative, neutral or positive. Table 2 gives an illustration for the possible outcomes of the statistical rule significance test (T) and indicates its

Table 2: **Interpretation of *Statistical Rule Significance***

Outcome	*Interpretation*
T <0	- Itemset X has a negative influence on the occurrences of itemset Y - Interest between 0 and 1 - Valid rule (T = -)
T is neutral	- Interest =1: X and Y are statistically independent → rule gives no extra information - Interest ≠1: rule has failed the χ^2- test → rule is not valid
T >0	- Itemset X has a positive influence on the occurrences of itemset Y - Interest >1 - Valid rule (T = +)

relation with the interest of the rule. 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 interest value and removing the rules that give no additional information towards this traffic accident analysis. An example of such a rule is:

Wet Þ *Rain* (sup= 19,91%, conf = 68,21%, , T = +, I= 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. Therefore, this rule will be removed from the association rules set.

3.3 **Results**

This paragraph 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 (section 3.3.1), collision in parallel (section 3.3.2), sideways collision (section 3.3.3), week/weekend accidents (section 3.3.4) and weather conditions (section 3.3.5). 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 interest value.

3.3.1 *Collision with a Pedestrian*

Table 3 illustrates that in 60,78% of all accidents involving pedestrians, collisions occur on crossroads 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 reach a maximum of two meter (3). This distance will probably 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 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 great number of children that head for school, and therefore will be on the Belgian roads, around these times.

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, collisions with pedestrians 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.

Table 3: **Rules for collisions with a pedestrian**

N	SUP	CONF	T	I	BODY		HEAD
1	8,65	62,75	+	7,25	[pedestrian]	=>	[visible]
2	6,76	49,02	+	7,25	[pedestrian]	=>	[zebra crossing with traffic lights]
3	5,59	40,52	+	7,25	[pedestrian]	=>	[unsheltered walking distance=2 meter]
4	5,50	39,87	+	3,16	[pedestrian]	=>	[one road user upwards, one transverse]
5	5,50	39,87	+	1,29	[pedestrian]	=>	[afternoon]
6	10,99	79,74	+	1,26	[pedestrian]	=>	[road with one roadway]
7	8,38	60,78	+	1,24	[pedestrian]	=>	[crossroad with traffic lights]
8	11,53	83,66	+	1,20	[pedestrian]	=>	[week]
9	10,00	72,55	+	1,19	[pedestrian]	=>	[daylight]
10	10,00	72,55	-	0,82	[pedestrian]	=>	[driving in a straight line]
11	8,11	58,82	-	0,75	[pedestrian]	=>	[constant speed]

3.3.2 *Collision in Parallel, Driving in the Same Direction*

Table 4 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 (1). 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 (2) and in 42,36% of the collisions in parallel, one of the road users will have used his brakes with the intention to stop (3). 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 (7) and not respecting the distance between different road users will often lead to more than one collision (6).

Finally, a collision in parallel will less often than expected coincide with a road user driving at a constant speed (5) and will have a smaller probability than expected of happening at a crossroad (4).

Table 4: **Rules collision in parallel**

N	SUP	CONF	T	I	BODY		HEAD
1	5,95	81,48	+	6,28	[distance]	=>	[parallel]
2	5,95	45,83	+	6,28	[parallel]	=>	[distance]
3	5,50	42,36	+	3,27	[parallel]	=>	[brake]
4	11,98	92,36	-	0,95	[parallel]	=>	[crossroad]
5	8,38	64,58	-	0,83	[parallel]	=>	[constant speed]
6	5,135	70,37	-	0,84	[distance]	=>	[one collision]
7	8,56	65,97	-	0,81	[parallel]	=>	[two people involved]

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.

3.3.3 *Sideways Collision*

The rules in table 5 indicate that when a sideways collision occurs, the road user will often not have respected the priority regulation of the crossroad (12). Most of the times he will also drive at a constant speed (15). These sideways collisions where the priority regulation of the crossroad is not respected, have a higher probability than expected of happening on crossroads where the road users should give way to the vehicles coming from the right (7).

Table 5: **Rules sideways collision**

N	SUP	CONF	T	I	BODY		HEAD
1	10,54	43,82	+	3,02	[no priority]+[crossroad with priority to the right]	=>	[local road]
2	14,00	75,24	+	2,48	[crossroad with traffic lights]+[no priority]	=>	[left turn]
3	7,84	64,44	+	2,12	[crossroad with traffic lights]+[one road user upwards, one opposite]	=>	[left turn]
4	5,59	53,91	+	1,99	[no priority]+[local road]+[crossroad with priority to the right]	=>	[equal road functions]
5	14,41	52,63	+	1,73	[crossroad with traffic lights]+[sideways]	=>	[left turn]
6	21,53	70,92	+	1,45	[left turn]	=>	[crossroad with traffic lights]
7	19,46	49,43	+	1,43	[sideways]+[no priority]	=>	[crossroad with priority to the right]
8	5,77	80,00	+	1,34	[traffic lights]+[night with public lighting]	=>	[sideways]
9	24,05	69,35	+	1,34	[crossroad with priority to the right]	=>	[no priority]
10	20,36	67,06	+	1,29	[left turn]	=>	[no priority]
11	39,37	75,87	+	1,27	[no priority]	=>	[sideways]
12	39,37	66,01	+	1,27	[sideways]	=>	[no priority]
13	17,75	74,34	+	1,25	[one road user upwards, one downwards]	=>	[sideways]
14	21,35	70,33	+	1,18	[left turn]	=>	[sideways]
15	50,63	84,89	+	1,09	[sideways]	=>	[constant speed]
16	6,22	49,29	-	0,83	[one road user upwards, one transverse]	=>	[sideways]
17	5,95	48,53	-	0,80	[traffic lights]+[sideways]	=>	[daylight]
18	18,6	37,87	-	0,73	[crossroad with traffic lights]	=>	[no priority]
19	5,05	38,89	-	0,65	[break]	=>	[sideways]

When the priority on the crossroad is regulated by traffic lights, the sideways collision will often occur when a road user makes a left turn (5).

These results refer to the relation between not respecting the priority regulation of the crossroad and the type of the priority regulation. An accident that occurs on a crossroad 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 (8). An accident that occurs on a crossroad with traffic lights will on the contrary less frequently coincide with not respecting this priority regulation (18). In 75,24% of the accidents where this violation does occur with traffic lights, a road user will also have made a left turn (2). Moreover, 70,92% of all accidents that occur when a road user makes a left turn, take place on a crossroad with traffic lights (6).

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 crossroads where at least one of the roads is local (1) or where both of the roads have a local character (4). 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 (14) and often when making a left turn, a priority violation will be the cause of the accident (10). Not respecting the priority regulation of the crossroad will lead in 75,87% of the accidents to a sideways collision (11).

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 (9). This type of collision near the traffic lights will less frequently occur at daylight (17). 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 (13). We would rather expect that this road situation would lead to a frontal collision. However, these accidents will occur on crossroads with traffic lights where the road users will drive in opposite directions and at least one of them will make a left turn (3).

Finally, when one of the road users uses his brakes with the intention to stop (19) or when one vehicle is driving upwards in the street and another vehicle is driving transversal on this direction (16) the accident will less frequently be a sideways collision.

In conclusion, there are two types of sideways collisions. The first type takes place at crossroads 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 crossroads 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.

3.3.4 *Week/Weekend Accidents*
Weekend: from Friday 21 h. - Monday 6 h.
Morning: 6-11h. ; afternoon: 12-16h. ; evening: 17-23h. ; night: 24-5h.

As shown in table 6, most accidents that occur at night, will take place during the weekend (1). Moreover, the accidents that take place in the weekend will more often than expected occur at night with public lighting (2) and will less frequently occur at daylight (11). Similarly, the accidents that happen on Sunday will have a higher probability than expected of occurring at night with public lighting (3), 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 daytrips or family excursions.

However, accidents that happen at night do less frequently than expected coincide with a driver whose physical condition is normal (10). 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 (4). In general, accidents that occur at daylight will most of the times take place during the week (7).

As mentioned earlier, a collision with a pedestrian will also often occur during the week (5). Even more, accidents that happen during the week will have a higher probability than expected of occurring in the afternoon (7). The number of accidents that take place in the afternoon is accordingly smaller during the weekend than during the week (9).

Finally, 78,98% of the accidents that occur on crossroads where the crossing street is an important local road, take place during the week (6).

Table 6: **Rules week/weekend accidents**

N	SUP	CONF	T	I	BODY		HEAD
1	7,38	58,57	+	1,92	[night]	⇒	[weekend]
2	14,59	47,93	+	1,45	[weekend]	⇒	[night with public lighting]
3	5,766	45,07	+	1,36	[Sunday]	⇒	[night with public lighting]
4	8,018	80,91	+	1,33	[week]+[one road user upwards, one transverse]	⇒	[daylight]
5	11,53	83,66	+	1,2	[pedestrian]	⇒	[week]
6	11,17	78,98	+	1,14	[crossing important local road]	⇒	[week]
7	46,76	76,89	+	1,11	[daylight]	⇒	[week]
8	23,6	76,61	+	1,1	[afternoon]	⇒	[week]
9	23,6	33,94	+	1,1	[week]	⇒	[afternoon]
10	10,72	85	-	0,92	[night]	⇒	[normal physical condition]
11	14,05	46,15	-	0,76	[weekend]	⇒	[daylight]

To summarize, most accidents that occur during the weekend will take place at night and have a higher probability of occurring with a driver whose physical condition is not normal. Accidents that happen in daylight will more often occur during the week.

3.3.5 *Weather Conditions*

Table 7 illustrates that accidents on a wet road surface will have a higher probability than expected of occurring at night with public lighting (2) and a smaller probability of occurring at daylight (10). Accordingly, accidents that happen at night with public lighting will coincide more frequently than expected with a wet road surface (3) and less frequently with a dry road surface (8)

Similarly, accidents that take place in the rain will have a higher probability of occurring at night with public lighting (1) and a smaller probability of occurring at daylight (11).

Furthermore, accidents that happen on a wet road surface (4), when it rains (6) or that occur at night with public lighting will less frequently coincide with a driver 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 (5).

Finally, when more than two people are lightly injured, the accident will less frequently than expected have occurred on a dry road surface (9).

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 (5).

Table 7: **Rules weather conditions**

N	SUP	CONF	T	I	BODY		HEAD
1	9,73	48,87	+	1,48	[rain]	=>	[night with public lighting]
2	13,15	45,06	+	1,36	[wet]	=>	[night with public lighting]
3	13,15	39,78	+	1,36	[night with public lighting]	=>	[wet]
4	25,77	88,27	-	0,96	[wet]	=>	[normal physical condition]
5	18,56	93,21	-	0,96	[rain]	=>	[no alcohol]
6	17,39	87,33	-	0,95	[rain]	=>	[normal physical condition]
7	28,65	86,65	-	0,94	[night with public lighting]	=>	[normal physical condition]
8	19,64	59,4	-	0,84	[night with public lighting]	=>	[dry]
9	5,135	57,58	-	0,82	[3 lightly injured people]	=>	[dry]
10	14,41	49,38	-	0,81	[wet]	=>	[daylight]
11	8,919	44,8	-	0,74	[rain]	=>	[daylight]

4 CONCLUSIONS AND FUTURE RESEARCH

In this paper, the technique of association rules was used on a large dataset of traffic accidents to profile black spots in terms of accident related data and location characteristics. 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. For example, 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 and priority to the right and making a left turn are the most important factors in sideways collisions .

Although the analysis carried out in this paper revealed several interesting rules 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 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. Secondly, considering the large number of attributes in the traffic accident dataset, it seems interesting to explore the potential of techniques that generate rules with long patterns to uncover more complex associations in traffic accidents. Finally, the strength of the association algorithm to identify patterns in the context of traffic accident data could be further examined by comparing the accident types that are discovered from the results with the accident groups that can be discerned with a clustering technique.

5 REFERENCES

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, 1993, pp. 207-216.

Agrawal, R., Mannila, H., Srikant, R. et al. (1996) Fast discovery of association rules, in Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. et al (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California, USA, pp. 307-328.

Belgian Institute for Traffic Safety (BIVV) and National Institute for Statistics (2000), Year Report on Traffic Safety 2000, BIVV v.z.w., Brussels.

Brin, S., Motwani, R. and C. Silverstein (1997) Beyond market baskets: generalizing association rules to correlations, *Proceedings of the ACM SIGMOD Conference*

on Management of Data, Tucson, Arizona, USA, May 13-15, 1997, pp. 265-276.

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

Frawley, W., Piatetsky-Shapiro, G., and C. Matheus (1991) Knowledge discovery in databases: an overview, in Frawley, W. and G. Piatetsky-Shapiro (eds.), *Knowledge Discovery in Databases*. AAAI Press/ MIT Press, Menlo Park, California, USA, pp. 1-27.

Holte, R.C. (1993) Very simple classification rules perform well on most commonly used datasets, *Machine Learning* **11**, pp. 63-90.

Kononov, J. and Janson B. (2002) Diagnostic Methodology for the detection of safety problems at intersections, *Proceedings of the Transportation Research Board*, Washington D.C., USA, January 13-17, 2002, Paper No. 02-2148.

Lee, C., Saccomanno, F. and B. Hellinga (2002) Analysis of Crash Precursors on Instrumented Freeways, *Proceedings of the Transportation Research Board*, Washington D.C., USA, January 13-17, 2002, Paper No. 02-3790.

Mannila, H. (1997) Methods and problems in data mining, *Proceedings of the International Conference on Database Theory*, Delphi, Greece, January 8-10, 1997, pp. 41-45.

Silverstein, C., Brin, S. and R. Motwani (1998) Beyond market baskets: generalizing association rules to dependence rules, *Data Mining and Knowledge Discovery* **2(1)**, pp. 39-68.