Clustering and Profiling Traffic Roads by means of Accident Data

K. Geurts, T. Brijs, G. Wets

PROMOTOR ▶ Prof. dr. G. Wets
ONDERZOEKSLIJN ▶ Kennis verkeersonveiligheid
ONDERZOEKSGROEP ▶ LUC DAM, LUC BMA, PHL, VUB, VITO
RAPPORTNUMMER ▶ RA-2003-27

UNIVERSITAIRE CAMPUS
GEBOUW D
B 3590 DIEPENBEEK
T ▶ 011 26 81 90
F ▶ 011 26 87 11
E ▶ info@steunpuntverkeersveiligheid.be
I ▶ www.steunpuntverkeersveiligheid.be

Clustering and Profiling Traffic Roads by means of Accident Data

RA-2003-27

K. Geurts, G. Wets, T. Brijs, K. Vanhoof

Onderzoekslijn Kennis verkeersonveiligheid

DIEPENBEEK, 2003.
STEUNPUNT VERKEERSVEILIGHEID BIJ STIJGENDE MOBILITEIT.

Documentbeschrijving

Rapportnummer:	RA-2003-27
Titel:	Clustering and Profiling Traffic Roads by means of Accident Data
Ondertitel:	
Auteur(s):	K. Geurts, G. Wets, T. Brijs, K. Vanhoof
Promotor:	Prof. dr. G. Wets
Onderzoekslijn:	Kennis verkeersonveiligheid
Partner:	Limburgs Universitair Centrum
Aantal pagina's:	28
Trefwoorden:	clustering, accident locations, profiling, accident risk, frequent item sets, Poisson, Common covariance
Projectnummer Steunpunt:	1.2
Projectinhoud:	Analyse en detectie van zwarte zones

Uitgave: Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit, december 2003.

Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit
Universitaire Campus
Gebouw D
B 3590 Diepenbeek

T 011 26 81 90
F 011 26 87 11
E info@steunpuntverkeersveiligheid.be
I www.steunpuntverkeersveiligheid.be

Samenvatting

In het eerste deel van dit onderzoek wordt 'model based clustering' toegepast om 19 centrale wegen in Hasselt in te delen in verschillende groepen op basis van hun gelijkaardige ongevalfrequenties in 3 opeenvolgende periodes van elk 3 jaar: 1992-1994, 1995-1997, 1998-2000. Er wordt verondersteld dat het geobserveerd aantal ongevallen afkomstig is van een aantal dichtheidsverdelingen waarbij de parameters van de verdelingen en het aantal verdelingen of clusters en de grootte van deze clusters onbekend zijn. Het doel van 'latent class clustering' is, gegeven de onderliggende data, deze afzonderlijke dichtheidsverdelingen en het aantal en de grootte van de clusters te identificeren. Meer bepaald maken we gebruik van een multivariaat Poisson mixture model met een gemeenschappelijke covariantieterm om de data te modelleren. Een algemeen algebraïsche modelleringsysteem wordt gebruikt om de loglikelihood functie te maximalizeren. De ongevallendata voor dit onderzoek zijn afkomstig van het Belgisch analyseformulier voor verkeersongevallen met lichamelijke letstel. Deze data bevatten een grote hoeveelheid aan informatie omtrent de omstandighden waarin deze ongevallen hebben plaats gevonden: verloop van het ongeval, verkeersgegevens, omgevingskarakteristieken, wegcondities, menselijke factoren en geografische kenmerken. In het tweede deel van dit onderzoek wordt dan ook gebruik gemaakt van de techniek van frequente sets om elke cluster van straten te profileren in termen van de bovengenoemde beschikbare ongevalsdata. De sterkte van deze data mining techniek is het identificeren van ongevalskarakteristieken.

Summary

In the first part of this research, model-based clustering is used to cluster 19 central roads of Hasselt into distinct groups based on their similar accident frequencies for 3 consecutive time periods of each 3 years: 1992-1994, 1995-1997, 1998-2000. The observed accident frequencies are assumed to originate from a mixture of density distributions for which the parameters of the distribution, the size and the number of clusters are unknown. It is the objective of latent class clustering to 'unmix' the distributions and to find the optimal parameters of the distributions and the number and size of the clusters, given the underlying data. More specifically, we use a multivariate Poisson mixture model with one common covariance term to model the data. A general algebraic modelling system is used to maximise the loglikelihood function. The accident data are obtained from the Belgian "Analysis Form for Traffic Accidents" and contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident, traffic conditions, environmental conditions, road conditions, human conditions and geographical conditions. In the second part of this paper, the data mining technique of association rules is used to profile each cluster of traffic roads in terms of the available traffic accident data. The strength of this approach lies within the identification of accident circumstances that frequently occur together for each group of traffic roads. This can, in turn, make a strong contribution towards a better understanding of the accident circumstances in these clusters.

Inhoudsopgave

1. INTRODUCTION

In Belgium, every year approximately 50.000 injury accidents occur in traffic, with almost 70.000 victims, of which 1.500 deaths (Belgian Institute for Traffic Safety, 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. In Belgium, this macro-economic loss due to the lack of traffic safety on the roads is estimated at 3.72 billion Euros per year (Dielemann, 2000). Accordingly, traffic safety is currently one of the highest priorities of the Belgian government.

Cameron (1997) indicates that clustering methods are an important tool when analyzing traffic accidents as these methods are able to identify groups of road users, vehicles and road clusters which would be suitable targets for countermeasures. More specifically, cluster analysis is a statistical technique that groups items together on the basis of similarities or dissimilarities (Anderberg, 1973). In Ng, Hung and Wong (2002) a combination of cluster analysis, regression analysis and Geographical Information System (GIS) techniques is used to group homogeneous accident data together, estimate the number of traffic accidents and assess the risk of traffic accidents in a study area. The results will help authorities effectively allocate resources to improve safety levels in those areas with high accident risk. In addition, the results will provide information for urban planners to develop a safer city.

Furthermore, according to Kononov (2002), it is not possible to develop effective countermeasures to improve traffic safety without being able to properly and systematically relate accident frequency and severity to a large number of variables such as traffic, geometric and environmental factors. Lee, Saccomanno and Hellinga (2002) indicate that in the past, statistical models have been widely used to analyze road crashes. However, Chen and Jovanis (2002) demonstrate that certain problems may arise when using classic statistical analysis on datasets with such large dimensions such as an exponential increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables. This is where data mining comes to play. Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from large amounts of data (Frawley et al, 1991). From a statistical perspective it can be viewed as a computer automated exploratory data analysis of (usually) large complex data sets (Friedman, 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 (Mannilla, 2000). 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 modelling. However, when dealing with a complex data set of road accidents, the use of data mining methods seems particularly useful to identify the relevant variables that make a strong contribution towards a better understanding of accident circumstances.

Therefore, in this research we will illustrate the possibility of identifying geographical locations with high accident risk by means of clustering techniques and profiling them in terms of accident related data by means of data mining techniques using a small but complex data set of traffic accidents. In particular, in the first part of this paper we will use latent class clustering (also called model-based clustering or finite mixture modelling) to cluster traffic roads into distinct groups based on their similar accident frequencies. In the second part of this paper, the data mining technique of frequent item sets is used to profile each cluster of traffic roads in terms of the available traffic accident data.

The remainder of this paper is organized as follows. First, an introduction to the clustering technique and the concept of frequent item sets is provided. This will be followed by a description of the data set. Next, the results of the empirical study are presented. The paper will be completed with a summary of the conclusions and directions for future research.

2. LATENT CLASS CLUSTERING

As mentioned in the introduction of this paper, in this research an innovative method based on latent class clustering (also called model-based clustering or finite mixture modelling) is used to cluster traffic roads into distinct groups based on their similar accident frequencies. More specifically, the observed accident frequencies are assumed to originate from a mixture of density distributions for which the parameters of the distribution, the size and the number of clusters are unknown. It is the objective of latent class clustering to 'unmix' the distributions and to find the optimal parameters of the distributions and the number and size of the clusters, given the underlying data (McLachlan and Peel, 2000).

2.1 Modelling Accident Rates with Poisson Distribution

Since we do not know exactly what causes traffic accidents to happen, the approach is based on the idea of modelling the accident frequency as a Poisson-distributed random variable Y. In general, the Poisson random variable $Y_i(t)$ represents the number of occurrences of a rare event in a time interval of length t and is therefore well suited for modelling the number of accidents at location i over a certain period of time t (Brijs et al, 2003).

This means that we are given a number of locations ($i= 1, \ldots, n$) on which the random variable Y_i (i.e. accident rate) is measured over a certain period of time (t), e.g. weeks, months or years. We assume the discrete random variable $Y_i(t)$ to be distributed Poisson, where $y_i = 0, 1, 2, \ldots$ and the rate parameter $\lambda t > 0$, i.e.

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

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

2.2 The Finite Mixture Specification

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

For instance, in the case of a two-cluster model, we assume that there are two support points. In other words, we assume there are two groups of locations:

- a group of roads of size p_1 whose latent accident parameter $\lambda=\theta_1$

- and a second group of roads of size $p_2=(1-p_1)$ whose average accident rate $\lambda=\theta_2$, where $0<p_j<1$, *and* $\sum_{j=1}^{k} p_j = 1$ are the mixing proportions with $k=2$. Note that the mixing proportion is the probability that a randomly selected observation belongs to the j-th cluster.

Consequently, the two cluster model can be formulated as:

$$P[Y_i(t)=y_i] = P[Y_i(t)=y_i \mid \text{group1}].\ P[\text{group1}] + P[Y_i(t)=y_i \mid \text{group2}].\ P[\text{group2}]$$

$$= \frac{(\theta_1 t)^{y_i}\, e^{-\theta_1 t}}{y_i!}.p_1 + \frac{(\theta_2 t)^{y_i}\, e^{-\theta_2 t}}{y_i!}.(1-p_1)$$

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

For the two cluster model, the loglikelihood function is then defined as:

$$\text{LL}(p_1, \theta_1, \theta_2 \mid \text{data}) = \sum_{i=1}^{n} \ln\left(p_1 \frac{(\theta_1 t)^{y_i}\, e^{-\theta_1 t}}{y_i!} + (1-p_1)\frac{(\theta_2 t)^{y_i}\, e^{-\theta_2 t}}{y_i!} \right)$$

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

2.3 Determining the Number of Clusters

In some applications of mixture models, there is sufficient *a priori* information for the number of clusters k in the mixture model to be specified with enough certainty. For instance, when the clusters correspond to externally existing groups. However, in this research, the number of clusters has to be inferred from the data, along with the parameters.

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

$$AIC = -2L_k + 2\, d_k$$

$$BIC = -2L_k + ln(n)\, d_k$$

$$CAIC = -2L_k + [ln(n)+1]\, d_k$$

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

3. FREQUENT ITEM SETS

3.1 Association Algorithm

In the second part of this study, an association algorithm is used to profile each cluster of traffic roads in terms of the variables available on the traffic accident form. This data mining technique was first introduced by Agrawal et al. (1993). It can be used to efficiently search for frequently co-occurring variables in large amounts of data. More specifically, the association algorithm produces frequent item sets describing underlying patterns in data. In contrast with predictive accident models, the strength of this algorithm lies within the identification of accident circumstances that frequently occur together (Geurts et al., 2003)). Informally, 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" itemsets.

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. 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.

3.2 Interesting Patterns

The association algorithm generates all item sets that have support higher than minsup. However, a large subset of the generated rules itemsets will be trivial and a filter is needed to post-process the discovered item sets. Two properties of the association algoritm 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):

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

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

Table 1: Interpretation of Lift

Outcome	**Interpretation**
+ ∞> *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*

Besides ranking the item sets on their lift value, we can use a second measure, i.e. the interestingness measure (I) to limit the accident patterns to only the discriminating or useful ones (Anand et al., 1997), Geurts et al., 2003)).

$$I = \frac{S(A,B)_2 - S(A,B)_1}{\max\{S(A,B)_2, S(A,B)_1\}}$$

This interestingness measure is based on the deviation in support values of the frequent item sets discovered for two different clusters. The nominator S_2 –$S1$ measures the difference in support for the accident characteristics in cluster 2 (S_2) and cluster 1 (S_1). The expression max $\{S_2, S_1\}$ is called the normalizing factor as it normalizes the interestingness measure onto the scale [-1,1].

3.3 Example

For example, consider the following accident data set containing 3 accidents:

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 variables should occur in at least 60% of all the accidents before they are 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:

s (Rain, Crossroad)= 3/4 (75%).

- Interestingness= s_1 (Rain, Crossroad) – s_2 (Rain, Crossroad) / max $\{s_1, s_2\}$

= ((2/3) – (3/4)) / (3/4) = - 0,11

<1 : This value is close to '0', indicating that although the item set (Rain, Crossroad) is very descriptive for both data sets (lift value>1), this item set is not very discriminating between the two data sets.

4. DATA

This study is based on a data set of traffic accidents obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the years 1992-2000. The data are collected by means of 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. These 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, physical characteristics, …). On average, 45 attributes are available for each accident in the data set. More specifically, this analysis will focus on 19 central roads in the city of Hasselt for 3 consecutive time periods of 3 years each:1992-1994, 1995-1997, 1998-2000. In total, 142 accidents are included in the analysis.

5. EMPIRICAL STUDY

5.1 Clustering Traffic Roads

5.1.1 3-Variate Poisson Distribution with Common Covariance

As explained in the previous section, the number of accidents on 19 (n = 19) similar roads in Hasselt (Belgium) are considered for 3 following time periods of each 3 years. The idea is to cluster the traffic roads in groups based on the similarities in the number of accidents that occurred on the roads during each time period.

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

$Y_1 = X_1 + X_{123}$ (period 1= 1992-1994)

$Y_2 = X_2 + X_{123}$ (period 2= 1995-1997)

$Y_3 = X_3 + X_{123}$ (period 3= 1998–2000)

with Y_i = the number of accidents on a traffic road in period i and all X's independent univariate Poisson distributions with respective parameters (λ_1, λ_2, λ_3, λ_{123}).

Since a large number of variables that influence the number of accidents in a certain time period will be time specific (e.g. traffic intensities), we will use one Poisson distribution for each time period to approximate the number of accidents in period i (X_i). Furthermore, it can easily be seen that the occurrence of accidents on a traffic road over several time periods may be related (e.g. due to bad infrastructure). Therefore, correlations between the observations in each cluster are allowed by identifying the parameter $\lambda_{123,}$ which can be considered as a covariance factor that measures the risk of the area common to all time periods (Karlis, 2000).

5.1.2 Mining the Algorithm

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

Figure 1 and figure 2 show the evolution of respectively the loglikelihood and the information criteria for different clusters (k=1, …,5) of the 3-variate Poisson Mixture Model with common Covariance.

These figures indicate the use of the goodness of fit measures to determine the number of clusters: although the loglikelihood of the model increases when the number of clusters increases, the information criteria will not choose the maximum possible clusters to cluster the data. Considering the model complexity, the AIC selects 3 clusters whereas the CAIC and the BIC select only 2 clusters. This difference can be explained by the fact that the AIC does not consider the size of the dataset, whereas the CAIC and the BIC do penalize for this factor. However, note that the difference between the AIC value for 2 clusters (219,8) and for 3 clusters (220,3) is very small.

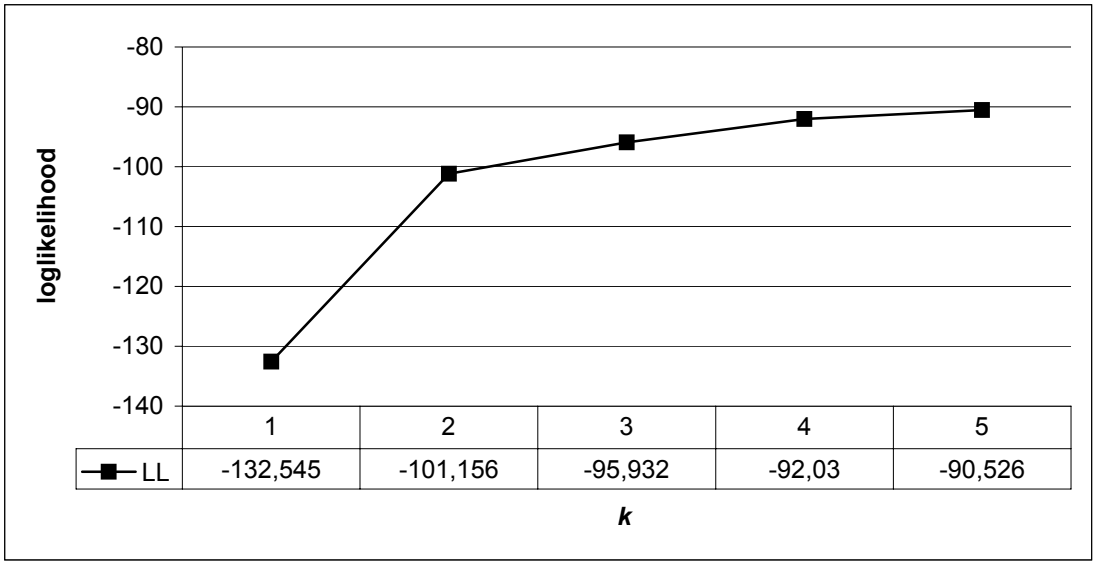

Figure 1: Loglikelihood against the number of clusters for the 3-variate Poisson Mixture Model with common Covariance

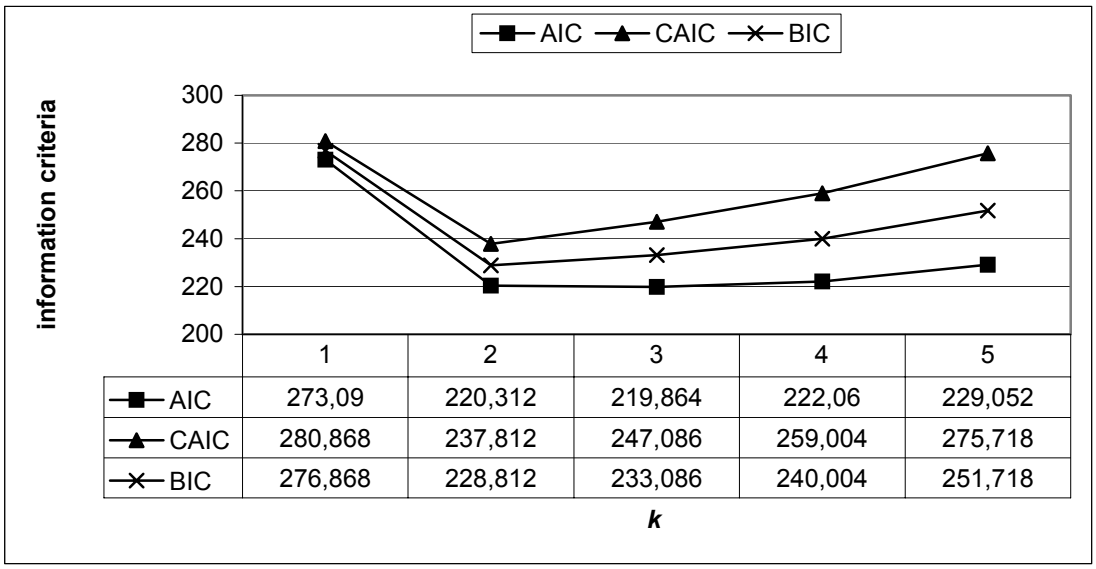

Figure 2: AIC, CAIC and BIC against the number of clusters for the 3-variate Poisson Mixture Model with common Covariance

5.1.3 Parameter Estimates

Table 1 and table 2 contain the parameter estimates and the size of each cluster (*p)* for the model with 2 and 3 clusters respectively. One can see that the cluster solutions are different.

In the 2-components common covariance model the average number of accidents increases per period for the first cluster and decreases per period for the second cluster. Furthermore, the observed average accident rate per period for cluster 1 is mainly dependent on the average accident frequency of the concerning period (λ_i) and less on the covariance factor (λ_{123}). For cluster 2, the covariance term does play an important role in the observed average accident rate per period. This can be explained as for this cluster there is a strong common factor in all periods that has to do with the accident risk on these roads, for example due to bad infrastructure, constant high traffic volume.

However, note that for this second cluster there is a strong decrease in the average number of accidents from period 2 (λ_2) to period 3 (λ_3). This could be an indication of infrastructural changes between these two periods.

Analogously, the results for the 3-components common covariance model can be analysed. One should remark that the value for λ_1 in cluster 1 and λ_2 in cluster 2 will be very small, meaning that the total average accident frequency for cluster1 in the first period and cluster 2 in the second year will mainly be influenced by the overall accident risk on the roads. Corresponding with the previous results, for cluster 3 a strong decrease in the average number of accidents from period 2 (λ_2) to period 3 (λ_3) can be found. This could be an indication of infrastructural changes between these two periods.

Table 1: Estimated parameters for the 2-components common covariance model

	Parameters				
Cluster	λ_1	λ_2	λ_3	λ_{123}	p
1	0,631	0,930	1,089	0,005	0,688
2	4,149	3,490	1,790	3,726	0,312

Table 2: Estimated parameters for the 3-components common covariance model

	Parameters				
Cluster	λ_1	λ_2	λ_3	λ_{123}	p
1	0,000	1,041	0,991	0,104	0,506
2	1,819	0,000	0,790	0,675	0,229
3	4,518	4,106	1,930	4,042	0,265

5.2 Profiling Traffic Roads

In the last part of this paper, we will use frequent item sets to profile each cluster of traffic roads. More specifically, we will focus on the results of the 2-components common covariance model which groups the traffic roads in two clusters. Since these clusters show different results for the overall accident 'risk' on the roads, one could expect that not every accident variable will be of equal importance when describing the different groups of traffic roads. Therefore, a comparative analysis between the accident characteristics that frequently occur together in the different clusters is conducted, which provides new insights into the complexity and causes of road accidents.

Two data sets of traffic accidents are defined according to the traffic roads belonging to cluster 1 and cluster 2. This is determined by estimating the posterior probability w_{ij}, i.e. the posterior probability for location i to belong to cluster j. This probability can be obtained for each observation vector y_i according to Bayes' rule. Indeed, after estimation, we know the density distribution f ($y_i \mid \theta_j$) with θ_j = vector of parameters for cluster j, and we know the cluster size p_j of each component such that we can calucalate the posterior distribution as:

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

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

…Cluster 1= Windmolenstraat; Jan van Helmontlaan; Vuurkruisenlaan; Zeven Septemberlaan; Weerstandslaan; Heldenplein; Nicolaas Cleynaertslaan; Helbeekplein; Kroonwinningsstraat; Elf Novemberlaan; Jan Palfijnlaan; Daniëlsstraat; Lentestraat

— Cluster 2= Kunstlaan; Casterstraat; Harpstraat; St. Katarinalaan; St. Katarinaplein ; Oude Luikerbaan

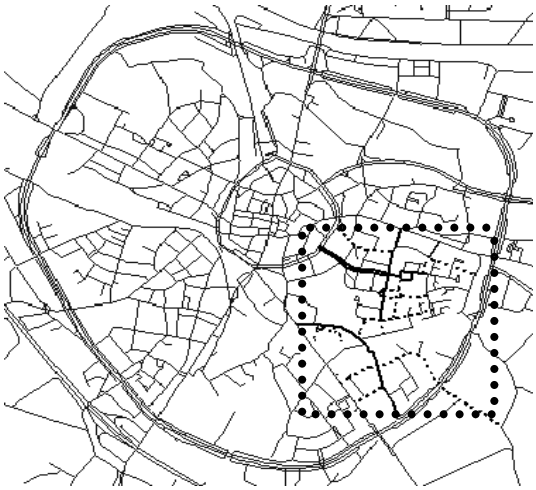

Figure 1: City map of Hasselt

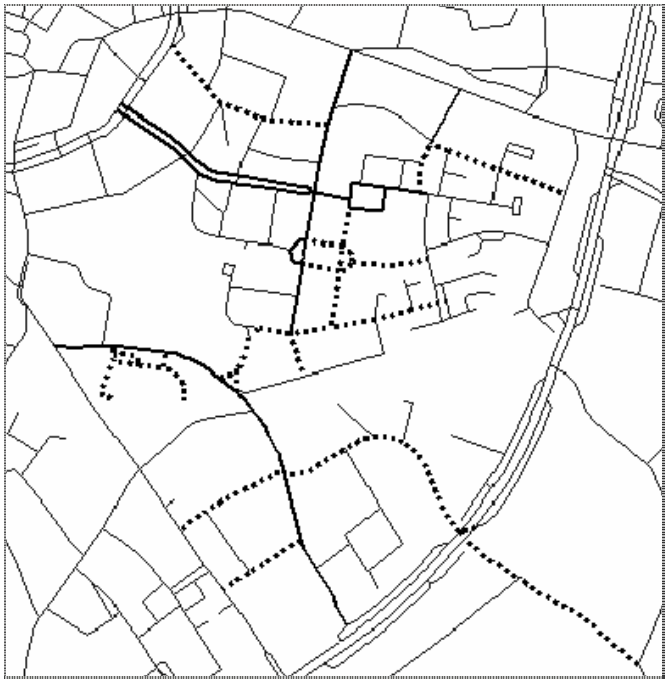

Figure 2: Traffic roads belonging to cluster 1 (…) and cluster 2 (—).

We 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.

5.2.1 Pre-processing and Transforming the Data Set

Some variables have a continuous character. Discretization of these continuous 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, 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 (7am-9am), morning (10am-12am), afternoon (1pm-3pm), evening rush hour (4pm-6pm), evening (7pm-9pm) and night (10pm-6am). A second example includes the variable age for which six new intervals were created: age between 0 and 17, age between 18 and 29, age between 30 and 45, age between 46 and 60 and age over 60. Other intervals were created by looking at the frequency tables for each variable. For variables where no domain knowledge for grouping the attribute values could be found, we used the *Equal Frequency Binning, a* discretization method to generate intervals containing an equal number of observations (Holte, 1993). Furthermore, attributes with nominal values had to be transformed into binary attribute values. This means that dummy variables were created by associating a binary 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.

A first analysis of these data sets shows that the two clusters of traffic roads have some common accident characteristics. For example, in both clusters most of the time 2 road users are involved in the accident (> 80%). Furthermore, most accidents occur on a crossroad (70-80%) with priority to the right (70-80%) where no priority was given (70-80%). Additionally, in both clusters most accident occurred in daylight (>80%), on a weekday (70-80%) with normal weather (> 80%) and on a dry road surface (70%).

However, some variables will occur more frequently in one cluster than in the other indicating differences between the accident characteristics on the roads of cluster 1 and cluster 2. For example, in cluster 2 the number of accidents involving a bicycle (18%) is almost twice as high as in cluster 1 (10%). Furthermore, in both clusters most accidents occurred in the morning (29%), however compared to cluster 1 in cluster 2 the number of accidents that occurred in the afternoon is relatively higher (24% compared to 17%) while the number of accidents that occurred in the morning rush hour is relatively lower (22% compared to 14%). Additionally, in cluster 2 the number of female road users involved in the accidents is slightly higher (44%) compared to cluster 1 (34%). Finally, in cluster 2 most road users are of the age between 18 and 29 (35%) and less of the other age categories while in cluster 1 most road users are of the age between 18 and 29 (33%) or 30 and 45 (33%).

These results already give an indication of the differences in accident characteristics between cluster 1 and cluster 2. In the following step, we will generate frequent item sets to identify combinations of accident characteristics that frequently occur together.

5.2.2 Generating Frequent Items Sets

A minimum support value of 30 percent was chosen for the analysis by means of frequent sets. It could be argued that the choice for the value of this 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 item sets. Accordingly, the number of rules that will be generated will cause further research on these results to be impossible due to computer memory limitations. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable to generate trivial rules.

From cluster 1, with a minsup=30 percent, the algorithm obtained 29415 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 support parameter the second analysis resulted for cluster 2 in 28541 frequent item sets of maximum size 4. These rules are further processed to select the most interesting rules.

5.2.3 Post-processing the Frequent Item Sets

As stated in the introduction of this paper, the emphasis in this part of the study lies on the profiling of clusters of traffic roads in terms of accident related data and the degree in which these accident characteristics are discriminating between the different clusters. Therefore, we will first discuss the item sets that are frequent for both clusters of accident locations. These accident patterns are descriptive for cluster 1 and for cluster 2. However, the occurrence of these patterns will not be equally strong in both data sets.

Selecting these frequent item sets resulted in 24562 accident patterns. The discriminating character of these accident patterns can be determined by means of the interestingness measure. In this research, we will pay special attention to the item sets with a positive interest value, i.e. approximating '1' since these accident patterns are stronger for cluster 2, i.e. the cluster with the highest accident risk.

Accordingly, selecting the item sets with $I > 0,3$ resulted in 12 item sets of size 2, 75 item sets of size 3 and 309 item sets of size 4. Table 3 gives an overview of the most interesting of these frequent item sets.

Table 3: Frequent Item Sets for Accidents in Cluster 1 and Cluster 2.

N	**Item1**	**Item2**	**Item3**	**Item4**	**S_2**	**$Lift_2$**	**S_1**	**$Lift_1$**	**I**
1	Weekday	Inside built up area	2 road users		71,96%	1,02	45,16%	0,94	0,37
2	Weekday	Age road user 18-29			59,81%	1,02	38,71%	0,92	0,35
3	Weekday	Inside built up area	Female road user		52,33%	1,01	32,25%	0,99	0,38
4	Weekday	Inside built up area	Female Road user	Car	54,20%	1,01	32,25%	0,75	0,40
5	Straight direction	50 km/h	Daylight		46,72%	1,06	32,25%	1,10	0,30
6	Sideways collision	50 km/h			48,59%	1,01	32,25%	1,05	0,33
7	Sideways collision	Inside built up area	Weekday	Car	57,94%	1,09	38,70%	1,03	0,33
8	Sideways collision	Inside built up area	2 road users	Weekday	57,94%	1,11	38,70%	1,20	0,33
9	Sideways collision	Female road user	Normal condition		55,14%	1,02	35,48%	0,88	0,35
10	Dry road surface	Inside built up area	Weekday		57,94%	0,99	38,70%	0,84	0,33
11	Normal weather	Inside built up area	Weekday	Straight direction	52,33%	0,97	32,25%	0,74	0,38

Results of table 3 show that the accident patterns that occur more frequently in cluster 2 than cluster 1 often occur on a weekday, inside the built up area with 2 road users [N=1], with one road user's age being between 18 and 29 [2]. Additionally, these accidents often involve a female road user [3], driving a car [4]. Furthermore, one road user is frequently driving in a straight direction with a speed limit of 50 kilometres per hour in daylight [5], These accidents frequently result in a sideways collision [6, 7, 8] with at least one road user in a normal condition [9].

The relatively young age (between 18 and 29) of the road users involved in the accidents on the traffic roads of cluster 2 could indicate that in this high risk cluster young parents are involved who drive their children to school. However, additional information on the traffic roads shows us that the schools are located on the traffic roads of cluster 1. On the traffic roads of cluster 2 the church, music academy, cultural centre and most important shops (e.g. bakeries, newspaper shops, hairdressers, laundry services, banks) are located. This information indicates that these accident patterns do not occur in the immediate environment of the schools but will most likely occur on the roads leading to and from the schools where the shops are located. Indeed, when looking at the time of accident, 14,5% of the accidents take place between 7am and 9am and 18,3% of the accidents take place between 4pm and 6pm. Furthermore, 29% of the accidents take place between 10am and 12am and 23,7% between 1pm and 3pm, which correspond with the opening hours of most shops. Remarkably less accidents occur in the evening (5,3% between 7pm and 9pm) and at night (9,2% between 10pm and 6am) when the shops are closed. Additionally, the accidents occurring on a weekday, inside the built up area and resulting in a sideways collsion mostly occurred on a crossroad (71,6%).

Furthermore, results of table 3 show that compared to cluster 1 the accidents on traffic roads belonging to cluster 2 occur more frequently on a dry road surface [10] and with normal weather [11]. However, note that for both clusters these accident patterns have a lift value smaller than '1'. This means that although in cluster 2 more accidents occur under normal weather conditions and on dry road surfaces than in cluster 1, these accident patterns still occur less frequently than expected for both clusters.

These accident patterns indicate that most accidents that occur on a traffic road belonging to the high risk cluster (cluster2) take place under no special variable environmental circumstances (e.g. rain, alcohol). Therefore, it can be expected that the high number of accidents on these traffic roads can be explained by an unsafe infrastructure or a high traffic volume for all time periods, confirming our previous results of a high common covariance factor for this cluster.

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

Selecting the item sets that are unique for cluster 2 resulted in 3943 frequent accident patterns. Table 4 gives an overview of the most interesting of these frequent item sets.

Table 4: Frequent Item Sets for Accidents in Cluster 2.

N	**Item1**	**Item2**	**Item3**	**Item4**	**Support**	**Lift**
12	Sideways collision	Female road user	No priority given		41,12%	1,23
13	Sideways collision	Female road user	No priority given	Crossroad	34,57%	1,58
14	50 km/h	Brakes	No priority given		30,84%	1,40
15	50 km/h	Car	Age road user between 18 and 29		39,25%	1,19
16	Weekday	Bicycle			33,64%	1,09
17	Weekday	Bicycle	2 road users		30,84%	1,14
18	0 deadly injured	Bicycle			36,44%	1,13

Conform with the results of table 3, results of table 4 show that sideway collisions involving female road users are a typical accident pattern for traffic roads with a high accident risk [12, 13]. Again, these results indicate that this type of accident occurs frequently while the maximum speed limit was 50 kilometres per hour for these accidents, while no priority is given [12, 13, 14] and the age of at least one road user was between 18 and 29 [15].

A second important accident type that is reflected in the results of table 4 are the accidents involving a bicycle. These accidents often take place on a weekday [16] with 2 road users [17] and frequently coincide with 0 deadly injured victims [18]. Note that these accident patterns are not very surprising as such, but remark that they do not appear for the accidents of cluster 1. Again, this could be explained by the proximity of shops, the cultural centre, music academy, church etc. on the traffic roads of cluster 2. The intensity of bicyclists will probably be much higher on these roads compared to majority of the roads of cluster 1 where in general none of these centres or stores are located. Since this traffic intensity of bicyclists will more or less be the same over all time periods, this factor will probably contribute to the high common covariance term for cluster 2.

However, since the schools are located on the traffic roads of cluster 1 and accordingly the intensity of bicyclists will also be high in these specific streets, it is surprising that these accident patterns with bicyclists do not occur at all in cluster 1. Again, this indicates that accidents with bicycles do not frequently occur in the immediate environment of the schools but will most likely occur on the roads leading to and from the schools.

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

Selecting the item sets that are unique for cluster 1 resulted in 4879 frequent accident patterns. Table 5 gives an overview of the most interesting of these frequent item sets.

Table 5: Frequent Item Sets for Accidents in Cluster 1.

N	**Item1**	**Item2**	**Item3**	**Item4**	**Support**	**Lift**
20	Crossroad	Priority to the right			61,61%	1,40
21	Crossroad	Priority to the right	Daylight		41,93	1,47
22	Normal weather	Priority to the right	Age road user between 30 and 45		32,25%	1,16
23	Normal weather	Dry road surface	Crossroad	Age road user between 46 and 60	35,48%	1,07
24	Car	Age road user between 46 and 60			35,48%	1,07
25	Inside built up area	weekend			32,25%	1,14

Results of table 5 show that an important accident type for the traffic roads with low accident risk are the accidents on crossroads with priority to the right [20, 21]. These accidents take up 61,61% of all accidents on these roads. However, in contrast with the previous results, these accidents more frequently than expected involve a road user with age between 30 and 45 [22]. Additionally, when an accident occurs on a crossroad with normal weather on a dry road surface, at least one road user of the age between 46 and 60 is involved [23, 24]. These results show that the age of the road user is not as pronounced for the accidents occurring on the low accident risk traffic roads.

Finally, an important accident pattern involves the accidents that occur inside the built up area in the weekend [25]. Note that these weekend accidents did not appear for the traffic roads with high accident risk.

6. CONCLUSIONS AND FURTHER RESEARCH

In the first part of this research, model-based clustering is used to cluster 19 central roads of Hasselt into distinct groups based on their similar accident frequencies for 3 consecutive time periods of each 3 years: 1992-1994, 1995-1997, 1998-2000. The strength of this technique lies in the identification of the optimal number of clusters and the size and parameters of each cluster. Results showed that the optimal number of clusters can vary from 2 to 3 clusters, depending on the chosen information criterion. For the two components model parameter estimates show that the average number of accidents increases per period for the first cluster and decreases per period for the second cluster. Furthermore, the observed average accident rate per period for cluster 1 is mainly dependent on the average accident frequency of the concerning period and less on the covariance factor. For cluster 2, the covariance term does play an important role in the observed average accident rate per period. This can be explained as for this cluster there is a strong common factor in all periods that has to do with the accident risk on these roads.

In the second part of this paper, the association algorithm was used on a data set of traffic accidents to profile the two clusters of traffic roads. The analysis showed that by generating frequent item sets the identification of accident circumstances that frequently occur together is facilitated. This leads to a strong contribution towards a better understanding of the occurrence of traffic accidents. This is particularly useful when dealing with a large and complex data set of traffic accidents of which the important variables. However, frequent item sets do describe the co-occurrence of accident circumstances but they do not give any explanation about the causality of these accident patterns. Therefore, their role is exploratory and to give direction to more profound research since the use of some additional techniques or expert knowledge will be required to identify the most important causes of these accident patterns, allowing governments to better adapt their traffic policies to the different kind of accident circumstances. Furthermore, the results indicate that the use of the association algorithm not only allows to give a descriptive analysis of accident patterns within one cluster, it also creates the possibility to find the accident characteristics that are discriminating between two groups of traffic roads.

The most important results indicate that sideway collisions on a weekday involving young road users are a typical accident pattern for traffic roads with a high accident risk. Furthermore, bicycle accidents are an important traffic safety problem in this cluster. These results could be explained by the proximity of shops, the cultural centre, music academy, church etc. and the resulting high traffic volume on these traffic roads. Accordingly, the roads belonging to this cluster should be considered as dangerous at all times resulting in a high number of accidents in all time periods. Furthermore, these accident patterns do not occur as frequently in cluster 1 although the schools are located in this cluster. For these traffic roads with a low accident risk crossroads with priority to the right are an important accident problem. However, in contrast with the results for the high accident risk traffic roads, these accidents occur in diverse age categories and also in the weekend. Therefore, we can conclude that on the traffic roads of cluster 1 most accidents occur by chance and less due to bad infrstructure. In conclusion, this analysis shows that a special traffic policy towards these clusters should be considered, since each cluster is characterized by specific accident circumstances, which require different measures to improve the traffic safety.

Although the association analysis carried out in this paper revealed several interesting patterns, which, in turn, provide valuable input for purposive government traffic safety actions, several issues remain for future research. First, the inclusion of domain

knowledge (e.g. traffic intensities, a priori infrastructure distributions) in the association algorithm would improve the mining capability of this data mining technique and would facilitate the post-processing of the association rules set to discover the most interesting accident patterns. Furthermore, the identified interesting accident patterns can be used in more statistical models to test their significance and to evaluate the difference in importance of these effects in the different clusters. Finally, the Poisson mixture regression model (Wedel et al., 1993) can be used to identify groups of locations with a different impact of the road characteristics on the accident risk. Indeed, it is possible that a specific combination of road characteristics is more dangerous for one group of locations while less dangerous for a second group of locations, depending on the other environmental factors that are not included in the analysis. The number of groups and the size of each group will be automatically identified by the technique and accordingly do not need to be defined in advance by the researcher.

ACKNOWLEDGEMENT

Work on this subject has been supported by grant given by the Flemish Government to the Flemish Research Center for Traffic Safety.

7. REFERENCES

Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases, **Proceedings** *of ACM SIGMOD Conference on Management of Data*, USA.

Agrawal, R., Mannila, H., Srikant, R., Toivonen H. and Verkamo A.I. (1996), Fast discovery of association rules, **Advances in Knowledge Discovery and Data Mining***,* 307-328.

Anderberg, M. R. (1973) *Cluster analysis for applications*, Academic Press, New York.

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

Brijs T., Karlis D., Van den Bossche F. and Wets G. (2003) A Bayesian model for ranking hazardous sites, Flemish Research Centre for Traffic Safety, Diepenbeek.

Cameron, A.C. and Trivedi P.K. (1986) Econometric models based on count data: comparisons and applications of some estimators and test, **Journal of Applied Econometrics, 1** 29-55.

Cameron, M. (1997) Accident Data Analysis to Develop Target Groups for Countermeasures. Monash University Accident Research Centre, Reports 46 & 47.

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

Dielemann L (2000) Huidige ontwikkelingen van het verkeersveiligheidbeleid (in Dutch), Belgian Institute for Traffic Safety (BIVV), Brussels.

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

Friedman, J. H. (1997) Data mining and statistics: What's the connection? **Proceedings** *of the 29th Symposium on the Interface Between Computer Science and Statistics,* Texas.

Geurts, K., Wets G., Brijs T. and K. Vanhoof (2003) Profiling high frequency accident locations using association rules, **Proceedings** *of the Transportation Research Board*, Washington D.C..

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

Karlis, D. (2000) An EM Algorithm for Multivariate Poisson Distribution and Related Models. Department of Statistics, Athens University of Economics and Business, Greece.

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.

Land, K. C., McCall, P. L., and Nagin, D. S. (1996) A comparison of Poisson, negative binomial, and semi-parametric mixed Poisson regression models—with empirical applications to criminal careers data. **Socio-logical Methods and Research, 24** (4), 387– 442.

Li, C-S, Lu, J-C, Park, J., Kim, K., Brinkley P.A. and Peterson J.P. (1999) Multivariate zero-inflated Poisson models and their applications, **American Statistical Association, 41** (1), 29-38

Lee, C., Saccomanno, F. and Hellinga B. (2002). Analysis of Crash Precursors on Instrumented Freeways, ***Proceedings*** *of the Transportation Research Board*, Washington D.C.

Manilla, H. (2000). Theoretical Frameworks for Data Mining. In *SIGKDD Explorations* 1 (2), 30-32.

McCullagh, P. and J.A. Nelder (1989) *Generalized linear models*, 2nd edition, Chapman and Hall, London.

McLachlan, G., and Peel, D (2000) *Finite Mixture Models*, Wiley Publications NY.

Ng, K-S, Hung, W-T and Wong W-G (2002) An algorithm for assessing the risk of traffic accidents, **Journal of Safety Research**, **33** 387-410.

Schwarz, G. (1978) Estimating the dimensions of a model, **The Annals of Statistics, 6** 461-464.

Wedel, M. Desarbo, W.S., Bult, J.R. and Ramaswamy V. (1993). A latent class poisson regression model for heterogeneous count data. **Journal of Applied Econometrics**, **8**, 397-411.