Road Traffic Accident Clustering With Categorical Attributes

Casaer Filip [a] Geert Wets [a *] Isabelle Thomas [b *]

a : Data Analysis & Modelling Group
Limburgs Universitair Centrum
Faculty of Applied Economics
Universiteits Campus – Diepenbeek 3590
Belgium.

b: Departement of Geography
Université Catholique de Louvain
Place Louis Pasteur 3
Louvain-la-Neuve 1348
Belgium

Fax : +32(0)11 26 87 00
Tel : +32(0)11 26 86 49
Tel : +32(0)10 47 21 36

filip.casaer@luc.ac.be
geert.wets@luc.ac.be
Isabelle@geog.ucl.ac.be

Total Number of words: 7577

[*] Corresponding authors

ABSTRACT

Road accidents are considered the result of a complex interplay between road user(s), vehicle(s), infrastructure and environment. To deal with this complexity and to disentangle – as much as possible – the attribute relationships, we here develop an unsupervised categorical model-based accident clustering. This technique enables us firstly to take the typical categorical data aspect into account. Secondly, instead of employing a heuristic measuring the distance between the accidents – as prevailing cluster techniques would do – it uses a more appropriate density-based similarity to assign the accidents to the different clusters. Finally, using all the available data unsupervisedly, the technique aims at an unbiased discovery of the data's inherent sub-structures. The method is applied to the road accident population observed in a Belgian suburban area (Brabant-Walloon). Our model partitioned this population into 5 clusters. Subsequently, all clusters were profiled, pointing out differences regarding time-dependency, type of road user(s), type of collisions, weather conditions, location, … . Since the determinative variables and the variable interplay clearly varied per clusters, they were studied accordingly. This accident examination at cluster level not only confirmed some existing findings but also generated new insights (or issues to get to the bottom of): the 'weekend accidents' are actually all-week accidents, the safety influence of passengers is subjected to weather conditions and the passenger formula, black zones consist mainly out of two accident types, confirmation of attribute relationships findings (e.g. age–gender) appears to be cluster dependent, ... Further research and knowledge discovery techniques can be applied within each of the clusters separately.

1. INTRODUCTION

In Belgium, every year approximately 50.000 injury accidents occur in traffic, with almost 70.000 victims, of which 1.500 deaths. The probability of having a deadly accident (relatively to the number of vehicle-kilometers traveled) was in 1998 in Belgium almost 35% higher than the European average. The European Road Statistics 2002(*1*) still displays Belgium's bad record towards traffic safety in comparison with most other European countries. 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 priorities of the Belgian government.

Road accidents – and the consequences of these accidents – are considered the result of a complex interplay between the driver, his vehicle and the road infrastructure. To deal with this complexity, we considered it useful to deploy an unsupervised traffic accident examination based on a broad clustering of all available attributes. A large number of studies have explored factors associated with accident occurrence and accident involvement. But each time these study focus on a specific sample of the accident population or choose out some specific contributing attributes. However, these studies also caution for the interaction between the attributes. For example when studying age and sex differences in the risk of causing vehicle collisions Claret et al (*2*) emphasized to account at least for the type of crash. Also Bédard et al (*3*), when examining accident contribution factors for single-vehicle crashes with fixed objects, such as age, gender, alcohol, vehicle speed, direction of impact, …etc. emphasized the interplay of these factors and the difficulty to control for the numerous confounding variables. These Canadian authors point out that further investigation of the relationships between the contributing variables is necessary. Vollrath et al.(*4*) examined the protective effect of carrying passengers. Different results showed up depending on age, infrastructure environment, daytime, traffic situations. Yau Kelvin (*5*) demonstrated that the risk factors affecting the severity of single vehicle traffic depended on the type of vehicle. Private vehicles were affected differently than goods transport or motorcyclists. Every important contributing variable seems to interact within the complexity of the traffic accident.

Congruently with the above findings and recommendations, we will examine the traffic data in this research without any assumption or hypothesis on either the existence of some typical in advanced known group or the relationship between contributing factors. A thorough knowledge discovery approach will show helpful to unravel the complicated interplay of the determining factors. Recent data mining research of Sohn (*6*) already demonstrated the usefulness of a cluster approach. To improve the classification accuracy for the severity of road accidents, the accident data was first clustered (clustering based on k-means algorithm) and subsequently a classification model was fitted for each cluster accordingly. The clustering enables a practical dimension reduction and an accident group demarcation. This approach may not only be more adapt to deal with the complexity of these harmful social events, but the resulting accident groups or clusters may also make up a useful input for further specific research. An other example of recent research based on a knowledge discovery methodology is Karlaftis' (*7*) hierarchical tree-based regression, assessing the effects of various highway geometric characteristics on accident rates.

The data we will deploy is derived from the National Institute for Statistics and represents a collection of the traffic accident records containing multivariate crash analysis features related to the accident type, the accident circumstances, the driver specifications and to other potentially influential factors. Furthermore we incorporated a geographic factor, i.e. a black zone attribute, in the modelling process to search for the connectivity between spatial an multivariate crash analysis.

The cluster algorithm we used is model-based. It matches our knowledge discovery approach and can deal with the categorical aspect of the traffic accident data. As proven by to McLachlan et (*8*) and Vermunt et al. (*9*), model-based (or latent class) clustering has gained a lot of attention during the last years and is considered a powerful methodology that provides a statistically sound basis for segmentation. The accidents, will be partitioned in a set of meaningful subclasses or clusters. Accidents, that are considered similar to one another, make up a natural grouping or structure and are said to belong to the same latent class.

The paper is organized as follows. First we will discuss the method used. We clarify why this model-based or latent class clustering best matches our unsupervised research approach of traffic accident data. The section contains also a formal description of the cluster technique. Secondly we will describe the data that we employed for the analysis. A substantial number of variables is needed when we try to compete against the complexity of traffic accidents. Next the results of the empirical study will be presented and deployed to profile the different discovered clusters. Finally several interesting features and attribute relationships, discovered while examining the cluster specific attribute value distributions – originating from the 32 variables and their categories – are discussed. The paper will be completed with some practical considerations, conclusions and indications for future research.

2. METHOD

As previously mentioned, we opted for a model-based or latent class cluster analysis. According to Kaufman and Rousseeuw (*10*)**,** cluster analysis is "the classification of similar objects into groups, where the number of groups, as well as their forms are unknown". This same definition could be used for our exploratory model-based analysis where a K-class latent variable is used to explain the associations among a set of observed variables. In the context of this research the k categories of the latent variable represent the different accident groups or clusters we are looking for. In the first part of this section we will confront you with the weak applicability of the traditional cluster techniques, compared with this paper's technique. Next, a formal description will clarify the relevant concepts for the development of the cluster model.

2.1 Cluster Techniques & Applicability

The more prevailing clustering techniques, as there are the partitioning techniques (e.g. K-means) or the hierarchical techniques (e.g. Ward) qualify less for the this paper's research for several reasons (see also Brijs et al (*11*)). Firstly, these techniques are distance-based, i.e. the assignment to a cluster is based on a distance measure or a similarity index. These measures or indices are heuristics which are not designed to compare categorical data. Besides, the hierarchical technique would result in a computational burden. Because all pair wise distances would have to be stored in an excessive distance matrix, implying large storage demands. A partitioning technique would be computationally more feasible, but the results heavily depend on the initial starting solutions being chosen and the method presumes uncorrelated data, a normal distribution of the attribute values, and spherical resulting clusters. K-means for example is a useful technique when the centroid (= the mean point of all points of the cluster) can be defined and a proper distance measure can be applied. Though it is not useful for our approach : we are dealing with categorical data, we do not want to make a prejudgement on the attribute distribution and we have a significant amount of correlation in our data.

Secondly, most of these traditional techniques require the number of clusters to be specified in advance, which is incompatible with our unsupervised knowledge discovery approach. Since the model-based clustering is based on a statistical approach, i.e. the observations are assumed to be generated from a mixture of underlying probability distributions, a number of statistical tests are available to check the validity of the model. So our method allows for a statistical treatment of model selection and helps to determine the number of accident clusters. As the goal or outcome is unknown in advance, our safety analyses methodology can provide innovatory insights into the complexity and causes of road accidents.

The recent increase in interest in latent class models is due to the development of extended computer algorithms, which allows today's computer to perform analysis on data containing more than just a few variables. In particular, the formalization of the EM algorithm has given a new impetus to the research of latent class models. However, a disadvantage we mention nevertheless is the lack of scalability. There are still restrictions both in terms of the number of records and the number of variables considered, which limits the utility in a domain where data is abundant.

Besides the better match of this model-based clustering technique with our categorical data and our unsupervised approach, the technique also brings along other advantages. Where traditional clustering techniques assign a subject to just one cluster, the assignment of clusters within the model-based clustering is carried out in a probabilistic way. McLachlan and Basford (*8*) claimed that LC clustering can be viewed as a probabilistic variant of K-Means clustering where probabilities are used to define "closeness" to each center. The clustering is sometimes mentioned in the context of fuzzy clustering, meaning that a subject may belong to different clusters according to some grade of membership. Furthermore, Bayes' rule can be used to classify unseen observations into the identified clusters, since their values on the indicator variables can directly be used to compute their individual posterior class-membership probabilities. Recent advances in model-based clustering (Vermunt and Magidson (*12*)) make up an other advantage, as they enable the inclusion of variables of mixed scale types (nominal, ordinal, continuous and count variables). Our accident attributes are predominantly nominal, but we also utilize ordinal and count variables**.**

Finally, It already has been pointed in the past that in the context of market research the model-based LC analysis proved to be a powerful tool to identify important market segments, e.g. Wedel and Kamakura (*13*).

2.2 Model-based clustering : formal description

The key idea in model-based clustering, also known as latent class clustering or finite mixture models, is that the observed data (in our case accidents) are assumed to originate from a mixture of density distributions for which the parameters of the distribution and the size and number of the segments are unknown. It is therefore the objective of this model-based clustering to unmix the distributions and to find the optimal parameters and the number and size of the segments, given the underlying data. Attribute values of accidents belonging to the same class are assumed to come from the same density distribution, whose parameters are unknown and have to be estimated. The following equation (*12*) shows $f(y_i)$, the unconditional density of the vector y, as a mixture of densities. p_k are the mixing proportions.

$$f(y_i) = \sum_{j=1}^{k} \quad p_k \text{ x } (f(y|\theta_j))$$

The population of interest consists of k subpopulations or latent classes. $f(y|\theta_j)$ stands for the density function for observation y from subpopulations j. The observation y is multi-dimensional. θ_j is a vector of parameters determining the subpopulations density. The interest lies in estimating these parameters and finding the values of the non-observable vector or latent variable which contains the cluster labels for each observation. In other words, the objective is to find the optimal values for the parameter vector (= finding $\theta_{optimal}$) such that the observations are more likely to have come from $f(y|\theta_{optimal})$ than from $f(y|\theta)$ for any other value of θ. The k classes or clusters represent k different groups with coherent attribute value distributions regarding simultaneously all observed variables.

On that account the log-likelihood (log *L*) of the above equation will be maximized following the maximum likelihood (ML) estimation approach. The software we applied (Latent Gold) uses for that reason both the expectation-maximization and the Newton-Raphson technique. To determine the number of segments unsupervisedly, a so-called information criterion is used to evaluate the quality of a cluster solution. Basically, information criteria are goodness of fit measures, which take model parsimony into account. The idea is that the increase of the likelihood of the mixture model on the dataset (which results in a better fit), is penalized by the increase in the number of parameters that was needed to increase the fit. The smaller the criterion, the better the model in comparison with another. So we will augment the number of clusters and therefore the fit until the criterion allows us to. The criteria we employed are the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC) :

$$\begin{aligned} BIC_{\log L} &= -2 \log L + (\log N) \text{ x npar} \\ AIC_{\log L} &= -2 \log L + 2 \text{ x npar} \\ CIAC_{\log L} &= -2 \log L + [(\log N) + 1] \text{ x npar} \end{aligned}$$

where Log*L* = log-likelihood
npar = number of parameters in the model
N = total number of cases

3. ACCIDENT DATA

In Belgium traffic accident data is being stored with the help of the National Institute of Statistics (NIS). Belgian police forces register the public road accidents, marked by death and injured casualties, in an “Analysis Form for Traffic Accidents”, which is considered to be a synthesis of the accident. It consists out of 26 sections concerning the accident or the involved road user. The traffic accident data contain a rich source of information: course of the accident (type of collision, direction of impact, type of 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, …). In total 32 attributes, mostly categorical, are available for each accident in the data set. Although there are some bias in the reporting and some necessary considerations on the data quality, the quantity of the information coded in the database and the number of accidents recorded allow us to apply data mining algorithms and/or extensive statistical research.

To cluster the traffic accidents, the nationally collected data first had to be pre-processed and accommodated. To deal with redundancy, missing values, the large amount of explanatory variables and the skewed character of the data, several variables went through a transformation. Some variables were discretisized into new categories. Some new – often count- – variables are an aggregation of different accident form rubrics or variables. Other variables (e.g. age) have been redivisioned into categories. After the transformations during this pre-processing, 32 variables (both categorical, ordinal and count) were retained for the cluster analysis. Table 1 lists the variables with their definition, data mode and the number of the categories. When marked with '*', the different categories can be found in table 2.

<INSERT TABLE 1 HERE>

<INSERT TABLE 2 HERE>

For this study the 1997-1999 accident records from a suburban area were deployed. Although the database goes back to 1991 we only choose to apply the cluster algorithm on a more consistent three year period. Belgium is a quite small country but densely built and mainly urbanized (10,3 millions inhabitants; 30,528 km²). However, large disparities exist within the country (see e.g. Mérenne et al (*20*)). Hence, our analysis is here limited to one administrative region: the Walloon Brabant, which is a province extending South of Brussels. It is 1.100 km² large and counts almost 350.000 inhabitants. It is mainly characterised by urban sprawl, but also by the existence of some former small market towns like Nivelles, Braine-l'Alleud, Wavre or Jodoigne . The Eastern part is still quite rural, the western part more industrial. Limiting the extend of the studied area enables one to better control for other sources of variations (mobility habits, friction of distance, mobility policies, etc).

<INSERT TABLE 3 HERE>

The 3-year period is long enough to limit the influence of random fluctuations and short enough to embank evolutions in externalities. Changes in road and traffic conditions become more limited: e.g., changes in traffic policy, consciousness-raising, changes in exposure and traffic volume , road constructions. Furthermore, we focused on the accidents of a certain geographic level, i.e. accidents that happened on the numbered roads. In Belgium this are the highways and the regional or provincial roads. They make up 11% of the road network but account for about half of the accidents. This pre-processing brought a significant advantage in return. This higher geographical level of accidents relishes - in Belgium - a higher localisation accuracy, which allowed us to add a spatial attribute, i.e. the black zone variable. This was defined by Flahaut et al (*14*) a high number of accidents at the hectometre concerned and a high number of accidents at its neighbouring hectometres). So police reported crash data could be combined with a geo-statistical accident concentration value. Finally, also multiple road user accidents were left out. The analysis is performed on accidents where maximum two road users were involved. In total 1.991 traffic accidents records were available for an extensive cluster analysis.

4. RESULTS

The unsupervised categorical clustering resulted in 5 separated accident groups (k=5). The entropy R-squared classification statistic amounted to 93% (see McLachlan and Basford (*8*)). These statistic indicates how well the model predicts class memberships. We attained the optimal model, with the best fit and the according optimal amount of clusters in several steps. The model's fit, measured by the log-likelihood, increased every time an extra cluster was added. The more clusters - and thus the more parameters - we add, the better the model will fit the data. Since the model fit will increase with every cluster we add, we also have to take some Information Criteria (IC) into account. The IC correct the fit for the parsimony (i.e. number of parameters) of the model. These criteria guard against over fitting. The Bayesian IC, the Akaike IC and the Consistent Akaike IC are the three information criteria we used. When comparing models, the lower value of these criteria, the better the model. When we added the sixth cluster the three IC stopped declining but increased. So the optimal model-based cluster result consists of 5 clusters. The subdivision of the accidents into these clusters is represented in Table 4.

<INSERT TABLE 4 HERE>

4.1 Cluster description

To identify the different accident groups and to point out their profiles, we checked for each accident subpopulation *when* the accidents tended to happen, which the typical *involved road users* were, *which type of accidents* typically occurred and *where* the accidents took place. To answers these questions and to allow for a consequently labelling of the distinct clusters, an enormous amount of cluster specific attribute value distributions – originating from the 32 variables with several categories – were thoroughly examined. The cluster characteristics present themselves by attribute value frequencies (=cp) which deviate from the population average (=pp). We used also notations cp1, cp2, pp1 and pp2 to refer to the frequencies when dealing with the first or the second road user of the population or subpopulation.

Cluster 1: Time independent single-vehicle black zone crashes on highways

The first cluster is the largest cluster containing 25,9 % of the 1.991 accidents, but only 21,7 % of the 3.270 involved road users. The cluster (like cluster 2) consists mostly of collisions with off-road obstacles after a loss of control or an unexpected evading. The cluster percentage (cp) amounts 59% for this type of accidents while the average population percentage (pp) is 30 %. Besides these single vehicle crashes the cluster contains many rear-end collisions (cp=25% while pp=18%) and almost no frontal collisions at all. The accidents took always place on road segments (pp=61%), always outside the built-up area (pp=65%), and congruently in the first place on highways (cp=83 % while pp=28%). 38% of the accidents happen within concentrated zones, labelled as black zones (pp= 29%). Furthermore, very typical for the first (and also the second) cluster, is their time-independency. This group of accidents took place both in the week and the weekend, at night or during the day, at rush hours or not at rush hours and no matter which season. This time-independency of the first cluster will be compared graphically with the time-dependency of the 5th cluster in the discussion section. First road users drove cars (cp for first road user or cp1= 89% while pp for first road user or pp1=82%). If present, the second road user drove besides cars also trucks (cp2=18% while pp=5%). The cluster presents a younger first road user population. The third age category (18-21 years, pp=7%) and the fourth (22-29 years, pp=15%) are relatively stronger presented with respectively 9 % and 21 %. Only 28 % of the second road users is female (pp2=33%). The cluster shows a characteristic presence of passengers simultaneously in front and at the back of the vehicle. This is the only cluster where this combination is more frequent than the presence of passengers only at the back.

Cluster 2: Time independent single-vehicle crashes on regional roads outside built-up area

The accidents of the second cluster consist for 45% out collisions with off-road obstacles (pp=30%). The accidents occurred again only on road segments, still 69% outside the built-up area, but this time only on regional roads. Within this cluster only 18% of the accidents took place at black zones (pp=29%). We find almost the same time-independence as in the first clusters, though these accidents occur less in the summer (cp=20% while pp=23%). According to the time-independence 37% of the accidents happened during the weekend (pp=32%). 83% of the first road users drove cars. This equals the population average. We find average presence of vans, trucks, motorbikes and mopeds. The second age category (15-17 years), the third (18-21 years) and the fourth (22-29 years) are more strongly represented among the first road users. Respectively 2% compared with 1% pp, 9% compared with 7% pp and 19% compared with 15% pp. Among the second road users we notice a cluster percentage of 17% concerning the 50-59 category (pp=11%). Only 25 % of the second road users is female (pp2=33%). Common infractions were: infractions on the right of way, place not in accordance, no distance kept. Furthermore the cluster accounts the highest registration of safety affecting personal conditions (e.g. alcohol, fatigue): 0,28 counts par person compared with a population average of 0,15.

Cluster 3: Lateral collisions on (mixed) crossroads on rainy week days with passengers involved

In this cluster accidents behave more time (or exposure) dependent. It contains only a small amount of single-vehicle crashes (12%). 57 % of the accidents were lateral collisions. The car is strongly present for both the first (89% while pp = 82%) and the second road user (84% while pp = 69%). All the accidents happened on crossroads mostly on regional ways (cp=91%), but also on highways exit or entries (cp=9%),. 30% of the crossroads showed difference on the maximum allowed speed between the roads (pp =20%). Common infractions were ignoring the right of way and ignoring the red light. 23% of the first road users was accelerating (pp=12%). The cluster has a very high passenger presence (45 % while pp = 25%) and is the most rainy cluster. It contains the highest percentage of accidents with precipitation, i.e. 60% (pp =41%).

Cluster 4: Lateral collisions with male two-wheelers involved

Also in this cluster accidents occur time or exposure dependently. Furthermore it exhibits a high afternoon accident frequency and a low weekend frequency (cp=22% and pp=32%). 53 % of the accidents are lateral collisions (pp 31%). The accidents took place on the regional roads, 60% occurred within the built-up area (pp=35%), mostly at crossroads (75% while pp=39%). It's a totally dry (98% while pp=61%) cluster with accidents occurring less in the winter (16% while pp=23%). The first road users of the cluster are only for 54% (pp=82%) made up by cars. In this group there is a significant presence motorbikes (16% while pp=4%), mopeds (15% while pp=4%) and bicycles (7% while pp=2%). Only 18% of them are women (pp=26%). The second road users are made up by the cars (86 % while pp=69%). Infractions on the right of way and the passing of a vehicle are the most common ones.

Cluster 5: Elderly (female) colliding with the vulnerable road users within the built-up area while accelerating

Cluster 5 is even more time or exposure dependent accident group then cluster 3 or 4. Accidents strongly occur around rush hours and 78% of the accidents occurred during the week (pp=68%). The accidents happened on the regional ways, 75% of them within the built-up area, both on crossroads (48%) and road segments. 36% of the accidents happened within black zones (pp=29%). The first road users drove cars (88%) while the second road users consisted almost entirely out of vulnerable road users : 18% motorbikes, 21% mopeds, 13% bicycles, 33% pedestrians. The first road users appeared to be rather older drivers: 40-49 category (pp1=18%) accounted 23%, 50-59 category (pp1=9%) accounted 10%, the 60-65 category (pp1= 3 %) made up 6% and the 65+ category (pp1=5%) made up 9%. Furthermore, 34% of the first road users were female (pp1=26%). This cluster has almost no passengers involved. 80% of these road users drove around without passengers (pp=70%). The second road users strongly represented the younger categories: 12% were under the age of 15 (pp2=2%), 16% was 15 or 16 years old (pp2=4%) and was aged in between 18-21 (pp2=9%). Besides the lateral collisions (45%) the cluster is mostly made up by pedestrian collisions (33%). Furthermore there are no single vehicle crashes present and 40% of the first road users is characterised by an acceleration dynamic.

4.2 Discussion

Due to the rather small amount of accidents involved, the clusters are not fully homogeneous. But while examining the enormous amount of cluster specific attribute value distributions – originating from the 32 variables contain several categories –several interesting features and attribute relationships were discovered. The list of topics we will discuss is not exhaustively. The point of view at cluster level is very interesting. Many existing findings can be checked for within this new context, which can result into a large amount of new findings. In this paper we are probably not covering all possible findings or future research trails that this new broad approach at cluster level can bring about. This section will try to show the possibilities of our approach.

Time dependent and time independent clusters.

The first two clusters are time-independent clusters. This is illustrated in figure 1 where the accident time distribution of the first two clusters is compared with the distribution of the fifth cluster.

<INSERT FIGURE 1 HERE>

Cluster 1&2 are clearly more leveled out over all moments of the day and the night. As describe previously also the week/weekend distribution behaved accordingly. These clusters contain clearly the well known phenomenon of the young male car drivers involved in single-vehicle crashes in the weekend. Apparently these accidents do not only happen within the weekends or at night. And considering also the heterogeneity still present in the clusters, these weekend accidents seem part of a bigger whole. They happen clearly all the time. They are independent of the amount of traffic and of the moment of the day or week. So when focussing on the phenomenon of the weekend accidents and it's typical targeted public, we should be aware of the totality of this safety problem. These type of accidents should perhaps rather be considered 'all-week accidents' instead of just 'weekend accidents'. Instead the time-dependent often lateral accidents could be considered 'week-accidents'. When focussing on these accidents we also point out that the first cluster is marked by a typical passenger formula (carrying passenger both in front and at the back) while the second cluster is marked by a higher influence of the personal condition (e.g. alcohol, fatigue).

Black zones, Clusters and Accident Severity

The first and the fifth cluster contain much higher percentages of black zone accidents (respectively 38% and 36% while pp=29%). As both clusters differ totally, this fact points at the existence of typical classes within the black zones. A future focus on black zone accidents will have to consider on the one hand the highway zones with high concentrations of single-vehicle crashes and on the other hand it will have to take the typical accidents with vulnerable road users on the regional roads within the built-up area into account. Further research and mapping of the accidents and the subpopulations will make a more thorough differentiation possible. Moreover, a geographical comparison of the accident groups and the whole of black zone accidents would be interesting to examine the calculation of accident concentration indices.

When considering the severity of the clusters, highway accidents are known to be more severe than accidents on other roads (confirmed by the European Statistical Report (*15*) on road accidents). Though when we examine our accident subpopulations and we compare the fatal injuries of the first cluster (highway) with the second cluster (regional road), we noticed an average of 0,029 deaths for each accident of the first cluster and 0,041 within the second cluster. So the most severe accidents do not seem to happen on highways but find themselves at the regional roads of cluster 2. Although the name of the zones could hint otherwise, the presence of black zones is clearly not associated with the severity of the clusters. The most lethal cluster showed the lowest black zone concentration (only 18%). This means that these accidents happen dispersed over the vast amount of regional roads within the study area, which makes the cluster a very hard target for policy makers. Nowadays investments in infrastructure works to deal with the black zones in our country will have little effect on this lethal accident group.

The safety effect of passengers.

When studying the influence of passengers Vollrath (*4*) pointed out that passengers increase total safety of the vehicle. Their social control on the drivers behaviour generally seemed to prevail on the distraction they can bring about. Our data and more particularly the registered infractions of the passenger cluster (cluster 3) accord with these findings. A loss of control (13%) and certain infractions as there are wrongly passing (2%), disregarding a safe distance in between cars (3%) occurred less frequently in this cluster (respective pp = 35%, 3%, 4%). On the other hand, ignoring a red light or giving no right of way – possibly caused by distraction – seemed to take place more frequently (respectively 4 times more and 2 times more).

The former research also pointed out that the effect was reversed when young male drivers were involved. We think there is not only an interaction of the safety effect with age and gender but also with passenger formula and/or weather conditions. Having passengers both in front and at the back of the car simultaneously is a typical feature present in the first cluster and more specific within this clusters weekend accidents (22 of the 35 accidents with these typical passenger formula took place in the weekend). So this type of passenger formula appears to be a dangerous one in some circumstances. When considering the combination of passenger presence and weather conditions, cluster 3 exhibits a certain attribute-relationship between passengers presence, precipitation and lateral collisions.

In accord with the findings of Vollrath et al (*4*) the positive effect of passengers is decreased at crossroads. This cluster, which only consists out of crossroad accidents, brings the hypothesis that the positive effect of passengers could be even more decreased at crossroads when it rains, causing lateral collisions.

Age, Gender and Type of Accident

The relevance of our cluster will be made clear through this topic by checking if some existing findings can be confirmed within our population and its clusters.

In the light of the rapid increase in the number of older drivers in our developing countries reflections on the oldest age categories gain more and more importance (see Maycock (*16*) . The results of the Claret et al (*2*) research in Spain suggest that the risk of causing a collision between vehicles with four or more wheels is directly dependent on drivers age. They concluded that for both sexes risk increased significantly with age. Regarding sex differences, they presented among young drivers higher risk ratios for men than for women When we accounted for the type of crash, as they suggested necessary for further research, our cluster model points out that these young male drivers are particularly involved in the first two single-vehicle crash clusters and rear end collisions on road segments. The (rather female) older drivers of cluster 5 are involved in lateral & pedestrian collisions within the built-up area. We had a closer look and crossed the age category with the type of accident. This confirmed the importance of the collision type when considering the influence of drivers age. Up to 60 % of the younger drivers categories were involved in single-vehicle crashes, while these only apply to 20% of the 65+ age category. These differences are even stronger within the population of female road users. While the middle-age road users are the ones strongly present within the rear end collisions, the elderly (65+) are more involved in lateral (40% of them) and frontal (19% of them) collisions. So to obtain a thorough view it appeared again necessary to account for a broader context e.g. the collision type. When examining the oldest age category within our cluster approach we note a high appearance in the third (30% of them) and the fifth cluster (25%). Claret (*2*) and Hakamies-Blomqvist (*17*) already suggested in the past that the aging feature is especially influential in complex situations when many stimuli must be processed. The profile of our 3rd and 5th cluster confirms the suggestions made : Cluster 3 accidents happen mostly during rainfall at crossroads with mixed maximum speed limits. Cluster 5 also represents accidents who took place at moments of multitasking within the built-up area. Furthermore, in the context of the fifth cluster the findings by that male gender is associated with particularly high risk of death among pedestrians are confirmed. A crossing of consequences and gender showed that 10 % of the male pedestrians died and only 5% of the female.

Our data does not confirm the Levine et al (*18*) findings concerning the interaction between age and gender: younger women would have an overall 50% increased risk for fatal injuries compared with men. The youngest fatally injured female (first) road user in our data contains belongs to the age category 30-39. The youngest fatally injured second road user belonged to the sixth cluster 50-59. A larger amount of records will be needed for a more thorough research.

Li et al (*19*) pointed out that female drivers were less likely to die in a car accident than male drivers, because women were involved in less serious accidents than men. The differences in fatal injury between man and women would be explained by behavioural differences between the sexes. Our data can confirm this : while 26% of the first road users were female, only 18% of the deceased first road users appeared to be female. But these results can differ strongly when examining the different clusters. None of the first road users of the fifth cluster, colliding with the vulnerable second road users within the built-up area died. 34% of this first road users were female (pp=26%). So this accident group helps confirming the above findings. These findings however are not confirmed in the first cluster : a proportional part of the fatally injured is female (28% while pp=29%). Female drivers here do not appear to behave differently. The different traffic subpopulations and accident clusters clearly have to be taken into account.

5. CONCLUSION

Our unsupervised model-based categorical clustering resulted in a subdivision of the 1997-1999 Brabant Walloon accident population into 5 different accident groups. The determinative variables and the variable interplay varied per clusters. Firstly each of the clusters became profiled, pointing out differences regarding time-dependency, type of road user(s), type of collisions, weather conditions, location, … . Secondly while examining the enormous amount of cluster specific attribute value distributions several interesting features and attribute relationships were discovered. It was not within the scoop of this paper to cover all possible findings or future research trails that this new broad approach at cluster level enables. Nevertheless several interesting remarks were made : The first two clusters showed a time-independent profile. The typical weekend-accidents which were involved appear to be all-week accidents. The most severe or fatally cluster is not located on the highways but on dispersed on the regional roads (it presented the lowest percentage of black zone accidents). Black zones seem to be made up out of single-vehicle crashes on the one hand and built-up area collisions of the (female) elderly with the young vulnerable road users. The safety effect of passengers is on the one hand confirmed, but interacts on the other hand not only with age and gender, but also with the passenger formula and with the precipitation. Furthermore our examination at cluster level showed to be useful in both clarifying and adjusting of findings on age, gender and their interaction.

Future research can be done within different directions. Firstly, investigating a larger amount of accident records will make a more homogenous clustering possible. Moreover, for further examination of the black zones it would be useful to develop a geographical mapping of the accident groups combined with these zones. Furthermore, a clustering per category, e.g. a clustering of every collision-type population separately, will help unravelling the complexity of the traffic accidents.

REFERENCES

(1) The Voice of the European Road : European Road Statistics 2002, European Union Road Federation, 2003.

(2) Claret P.L. et al., 2003, Age en sex differences in the risk of causing vehicle collisions in Spain, 1990 to 1999. Accident Analyses and Prevention 35, 261-272.

(3) Bédard, M. Guyatt, G.H. Stones M.J.,and Hirdes J.P. (2002) The independent contribution of driver, crash and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34, 717-727.

(4) Vollrath M. et al., 2002, How the presence of passengers influences the risk of a collision with another vehicle. Accident Analyses and Prevention 34, 649-654.

(5) Yau Kelvin K.W. , 2003 , Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analyses and Prevention, Article in Press (2003).

(6) Sohn, S.Y. Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accident in Korea, 2003.

(7) Karlaftis M.G., Golias I. (2002) Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis and Prevention 34, 357- 365.

(8) McLachlan, G., and Basford, K. E. Mixture Models. Marcel Dekker, INC, *New York Basel*, 1988

(9) Vermunt, J.K. & Magidson, J. Latent Class Cluster Analysis, chapter 3 in J.A. Hagenaars and McCutcheon (eds.) Advances in Latent Class Analysis. Cambridge University Press, 2000.

(10) Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data : An Introduction to Cluster Analysis, 1990.

(11) Brijs T., Swinnen G., and Vanhoof K. Retail Market Basket Analysis: A Quantitative Modelling Approach, Phd dissertation, 2002.

(12) Vermunt, J.K., and Magidson, J. Latent Gold 2.0 User's Guide, Belmont, MA : Statistical Innovations, Inc, 2000.

(13) Wedel , M. and Kamakura, W.A., A Simulated Likelihood Latent Variable Approach For Incomplete Mixed Outcome Data, 1998.

(14) Flahaut B, Thomas I., Mouchart, M. San Martin, E. The Local Spatial Autocorrelation And The Kernel Method For Identifying Black Zones : A comparative approach. Accident Analysis and Prevention 918, Article in Press, 2003.

(15) European Conference of Ministers of Transport : Statistical Report On Road Accidents 1997-1998, OECD Publications Service, 2001.

(16) Maycock, G., 1997. The Safety of Older Car-drivers in the European Union. European Road Safety Federation, ERSF, AA Foundation for Road Safety Research, Basingstoke, UK.

(17) Hakamies-Blomqvist, L., Older Drivers' Accident Risk : Conceptual And Methodological Issues, Acc. Anal. Prev. 30, 293-297 1998.

(18) Levine, E., Bédard, M., Molloy, D.W. Basilevsky, A., Determinants of driver fatality risk in front impact fixed object collisions. Mature Medicine Canada 2, 239-242. 1999.

(19) Li, G., Baker, S.P., Langlois, S.A., Kelen, G.D. Are Female Drivers Safer ? An Application Of The Decomposition Method, Epidemiology 9, 379-384. 1998.

(20) Mérenne, B., Van der Haegen H., Van Hecke, E., La Belgique - Diversité territoriale, in: *Bulletin du Crédit Communal*, n° 202, 1997/4.

LIST OF TABLES AND FIGURES

Tables & Graphs

Label	Definition	Mode	Categories
Id	Unique identification of accident	Continuous	/
Weekend	Time specification	Categorical :Nominal	No /Yes
Hour	Time specification	Categorical : Nominal	0 - 23
Season	Time specification	Categorical : Nominal	4 categories*
AccType	Describing direction of impact between road users, collision with obstacle, collision with pedestrian	Categorical :Nominal	8 categories*
Passenger position	Describing the position of the passenger in the vehicle	Categorical :Nominal	4 categories*
CrossRoadchar	Location specifications on priority regulation	Categorical :Nominal	4 categories*
Built-up Area	Area specification	Categorical :Nominal	Yes/No
Type	Road specification	Categorical :Nominal	Highway / Regional way
Soort	Road specification : Separation of lanes	Categorical :Nominal	Separated or not
Loc snelheid	Max allowed speed on road	Categorical :Ordinal	30 /50/ 60/ 90/ 120
Loc delta	Difference in Max allowed speed	Categorical : Nominal	Yes/ No
Black zone	Area specification	Categorical :Nominal	Yes /No
VisibilityAggr	Accounts for factors influencing the visibility	Count	/
MomentAggr	Accounts for moment specific factors	Count	/
StructuralAggr	Accounts for infrastructural factors	Count	/
PersonalAggr	Accounts for personal factors	Count	/
Detrimcounts	Weighted sum of human detriment in the accident	Count	/
First Road User			
Sort1	Type of first road user	Categorical :Nominal	9 categories*
Gend1	Gender of first road user	Categorical :Nominal	Male/Female
Age1	Age of first road user	Categorical: Ordinal	8 categories*
NumbPass1	Number of passengers	Count	/
Consequences1	Consequences for first road user	Categorical :Nominal	5 categories*
Behavior1 (motion or infraction)	Situational factor describing Motion & Positioning & Infraction	Categorical :Nominal	10 categories*
Dynamics1	Driving dynamics of first road user	Categorical :Nominal	5 categories*
Second Road User			
Sort2	Type of second road user (if there is one, otherwise category '0')	Categorical :Nominal	9 categories*
Gend2	Gender of second road user	Categorical :Nominal	Male/Female
Age2	Age of second road user	Categorical: Ordinal	8 categories*
NumbPass2	Number of passengers	Count	/
Consequences2	Consequences for second road user	Categorical :Nominal	5 categories*
Behavior2	Situational factor describing Motion & Positioning & Infraction	Categorical :Nominal	10 categories*
Dynamics2	Driving dynamics of second road user	Categorical :Nominal	5 categories*

Table 1: Description of the variables

Label	Categories									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
AccType		(Multiple)	Frontal	Rear end	Lateral	Pedestrian	Obstacle on the road	Obstacle off-road	1 road user, no obstacle	Unknown
Sort		Car	Van or Small Truck	Truck	Bus or Coach	Motorbike	Moped	Bicycle	Pedestrian	Other (**10 :** unknown)
Gender		Male	Female							
Age	0-15	16-17	18-21	22- 29	30-39	40-49	50-59	60-65	65+	
Consequences		Fatal	Heavily injured	Lightly injured	Uninjured	Fatal (30 days)				
Dynamics		Constant Speed	Brings to a stop	Accelerates	Stands still	Unknown				
Behavior (motion or Infraction)		Ignores red light	Gives no right of way	Crosses a continuous line	Passes wrongly	Evades unexpectedly	Place not in accordance	Loss of control	Keeps no distance	Continuous normal direction (**10** : road side)
Weather		Normal	Rainfall	Fog & Smoke	Wind	Snow- & hailstorm				
Crossroad		Crossroad with Traffic light	Crossroad with priority of main road	Other crossroad (Agent, right of way, flashing light)	No crossroad, but road segment					
Season		Winter	Spring	Summer	Autumn					
Passenger Position	No passenger	At the front seat	At the back seat	At both front and back seat	Not known where					
Type of road		Highway	Regional way	Regional way, in built up area						
Soort		1 roadway	2 roadways Separated roadway	Combination						

Table 2 : Description of different categories

1997-1999	Number of accidents	Involved road users	Lightly injured	Heavy injured	Death
België :	152794	292611	178632	32750	4261
Flanders	100932	196307	118054	20898	2412
Brussels	8770	18050	10675	683	128
Wallonia	43092	78254	49903	11169	1721

Table 3 : The number of Belgian accidents, with the involved road users and their consequences for the 1997-1999 period and the nations three regions.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total Population
Accidents (= the unit of analysis)						
Distribution of accident population	25,9 %	24,5 %	22,5 %	14,7 %	12,5%	100 %
Number of accidents	515	488	447	292	249	1.991
Road users						
Number of road users	711 (21,7%)	696 (21,3%)	833 (25,5%)	538 (16,5%)	492 (15,0%)	3.270

Table 4: Table: Distribution of the 1.991 accidents and 3.270 road users over the 5 different clusters

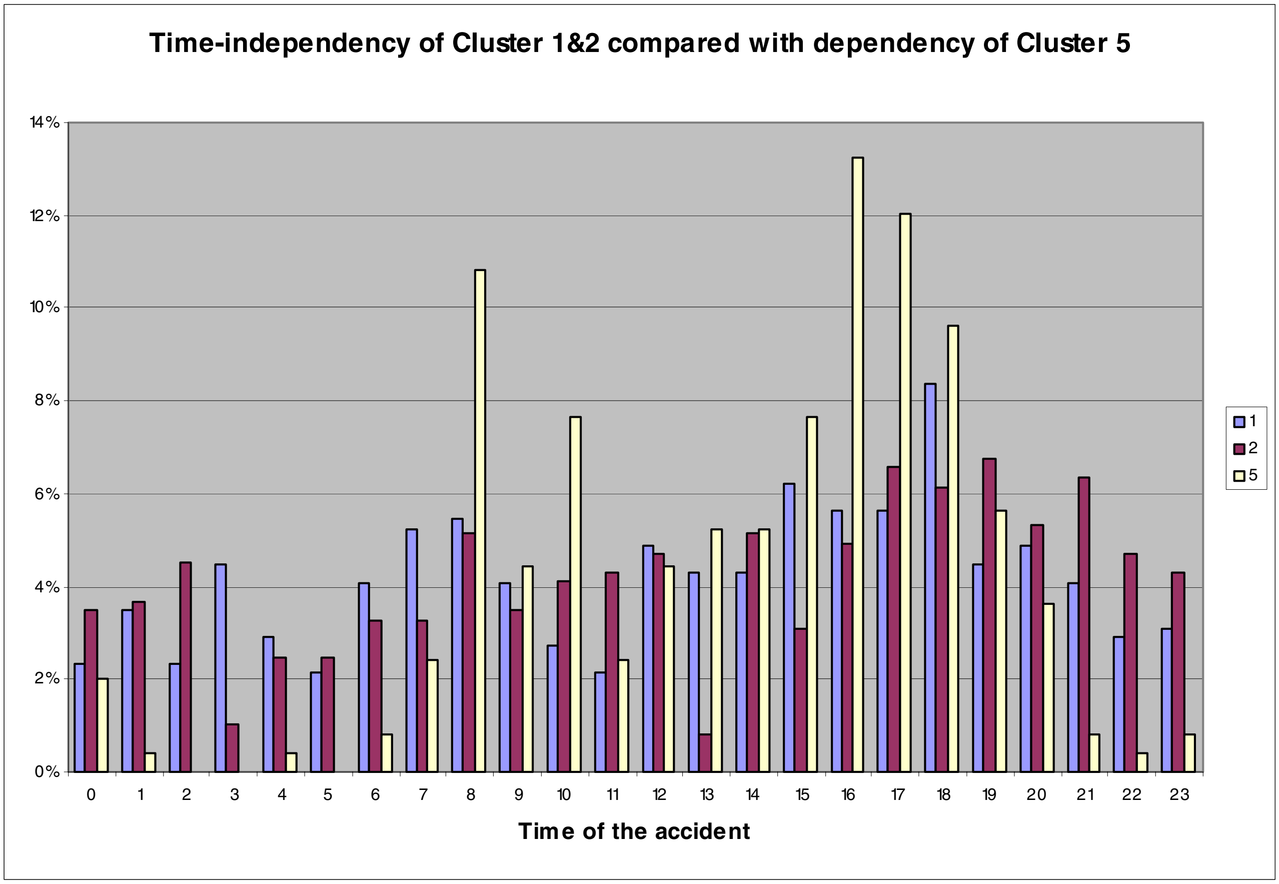

Figure 1: Accident time distribution : comparison between two time-independent (clusters 1&2) and a time-dependent cluster (cluster 5).