Traffic accident segmentation by means of latent class clustering

DEPAIRE, Benoit; WETS, Geert & VANHOOF, Koen (2008) Traffic accident segmentation by means of latent class clustering. In: Accident Analysis and Prevention, 40(4). p. 1257-1266.

DOI: 10.1016/j.aap.2008.01.007
Handle: http://hdl.handle.net/1942/8396

Traffic Accident Segmentation by Means of Latent Class Clustering

Benoît Depaire[(a)] [*benoit.depaire@uhasselt.be*]
Geert Wets[(a,b)] [*geert.wets@uhasselt.be*]
Koen Vanhoof[(a)] [*koen.vanhoof@uhasselt.be*]

(a): Transportation Research Institute
Hasselt University
Wetenschapspark 5 bus 6
3590 Diepenbeek
Belgium

(b): Corresponding author
Tel: +32 11 26 91 11
Fax: +32 11 26 91 99

Abstract

Traffic accident data are often heterogeneous, which can cause certain relationships to remain hidden. Therefore, traffic accident analysis is often performed on a small subset of traffic accidents or several models are built for various traffic accident types. In this paper, we examine the effectiveness of a clustering technique, i.e. latent class clustering, for identifying homogenous traffic accident types. Firstly, a heterogeneous traffic accident data set is segmented in seven clusters, which are translated into seven traffic accident types. Secondly, injury analysis is performed for each cluster. The results of these cluster-based analyses are compared with the results of a full-data analysis. This shows that applying latent class clustering as a preliminary analysis can reveal hidden relationships and can help the domain expert or traffic safety researcher to segment traffic accidents.

Keywords: Latent class clustering; Accidents; Heterogeneity; Injury analysis

1. Introduction

In order to improve traffic safety, traffic accidents need to be analyzed to identify possible risk factors and their effects on injury severity levels. Because in general traffic accident data are heterogeneous, researchers often try to reduce this heterogeneity. A common approach is to focus on a very specific traffic accident type. Examples are Bédard et al. (2002) who focus on single-vehicle crashes with fixed objects in their analysis, Zhang et al. (2000) investigating older drivers involved in injury motor vehicle crashes on public roads, Zajac and Ivan's research (2003) which focuses on pedestrians and several other studies which focus on a specific vehicle type such as sport utility vehicles (Ulfarsson and Mannering, 2004) or motorcycles (Shankar and Mannering, 1996; Quddus et al., 2002).

Other authors build separate models per traffic accident type (Valent et al., 2002; Yau, 2004; Chang and Mannering, 1999; Lee and Mannering; 2000; Ulfarsson and Mannering, 2004; Islam and Mannering, 2006; Carson and Mannering, 2001; Savolainen and Mannering, 2007). Analyzing the results of these separate models, we can discern three different situations where data heterogeneity can produce incorrect conclusions. Firstly, data heterogeneity can cause certain accident factors to remain hidden. Valent et al. (2002) found that Sundays and holidays have an increased injury risk for truck accidents. However, analyzing all traffic accidents together, Sundays and holidays remained hidden as an increased risk factor. Furthermore, Yau (2004) confirmed that the effects of risk factors may be obscured when combining the accident data from various vehicle categories and suggested to segment traffic accident data on vehicle type. Secondly, building separate models for different traffic accident types sometimes show how the effect of a risk factor on the injury outcome differs in magnitude between different traffic accident types. Islam and Mannering (2006) found that "driving without restraints, falling asleep, and overturned/rollover all resulted in an increased likelihood of injury for older females – more so than their male counterparts". Ulfarsson and

Mannering (2004) concluded in their work that there are significant differences in magnitude between males and females with regard to how various factors affect injury severity. Thirdly, factors affecting injury severity can differ in direction between different traffic accident types. Ulfarsson and Mannering (2004) found that male drivers of vehicles striking a barrier or guardrail experienced an increase in probability of less severity while female drivers experienced an increase in probability of greater severity. Sometimes building separate models for different traffic accident types shows that certain risk factors are not statistically significant for all traffic accident types. However, as Ulfarsson and Mannering (2004) mentioned, an insignificant variable in one model can simply be caused by a lack of observations. Therefore, one has to be careful to conclude that a factor has no significant influence on the injury outcome for a specific traffic accident type.

Most often, the segmentation of traffic accident data is based on expert domain knowledge, methodological decisions or the wish to study a specific problem. Although expert knowledge can lead to a workable segmentation of traffic accident data, it does not guarantee that each segment consists of a homogenous group of traffic accidents. Therefore, traffic accident analysis could benefit from a data analysis technique which aids in the process of traffic accident segmentation.

Such techniques can be found within the domain of data mining, which can be defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data (Fayyad et al., 1996). From a statistical point of view, data mining can also be considered as a computer automated exploratory data analysis of (usually) large complex data sets (Friedman, 1997). However, in contrast with statistics, data mining pays less attention to the large-scale asymptotic properties of its inferences and more to the general philosophy of “learning”, including consideration of the complexity of models and the computations they require (Hosking et al., 1997). Recently,

several data mining techniques have found their way into traffic safety research, such as rule induction (Kavsek et al., 2002), frequent item sets (Geurts et al., 2005), artificial neural networks (Mussone et al., 1999) and CART, i.e. classification and regression trees (Chang and Wang, 2006).

Depending on the objectives of the research, two major categories of data mining can be discerned (Berry and Linoff, 1997): predictive and descriptive techniques. For traffic accident segmentation, we opted for the descriptive data mining technique of cluster analysis, which divides heterogeneous data into several homogenous classes or clusters. The objective of this study is to examine the effectiveness of cluster analysis as a technique for identifying homogenous traffic accident types and to evaluate if it allows subsequent traffic accident analysis to reveal new information.

2. Methodology

Cluster analysis seeks to separate data elements into groups or clusters such that both the homogeneity of elements within the clusters and the heterogeneity between clusters are maximized (Hair et al., 1998). This technique is an unsupervised learning algorithm because the true number of clusters as well as their form are unknown (Vermunt and Magidson, 2002). As Xu and Wunsch (2005) mention in their review paper on clustering algorithms, cluster analysis has been applied in a wide variety of fields, ranging from engineering (machine learning, artificial intelligence, pattern recognition, mechanical engineering, electrical engineering), computer sciences (web mining, spatial database analysis, textual document collection, image segmentation), life and medical sciences (genetics, biology, microbiology, paleontology, psychiatry, clinic, pathology), to earth sciences (geography, geology, remote sensing), social sciences (sociology, psychology, archeology, education), and economics

(marketing, business) (Everitt et al., 2001, Hartigan, 1975, Green, 2004; Arabie, 1994; Moustaki and Papageorgiou, 2005; Jiang et al., 2004).

In its most common form, cluster analysis is based on heuristics which try to maximize the similarity between in-cluster elements and the dissimilarity between inter-cluster elements (Fraley and Raftery, 2002). These similarity-based clustering techniques use a specific distance function for elements with continuous features and similarity measures for elements with qualitative features. For elements consisting of both continuous and qualitative features, a mapping into the interval (0,1) can be applied such that a distance measure can be used. Among the similarity-based techniques, two major approaches can be discerned, i.e. the hierarchical approach (e.g. Ward's method, single linkage method) and the partitional approach (e.g. K-means). Although extensive research has been done in this field of heuristic-based cluster analysis, the statistical properties of these methods are generally unknown (Fraley and Raftery, 2002), whereas the statistical properties of probability model based clustering techniques (Bock, 1996; Fraley and Raftery, 2002) are better understood.

This second type of clustering is sometimes called mixture densities-based clustering (Xu and Wunsch, 2005), finite mixture models (Fraley and Raftery, 2002) or latent class models (Moustaki and Papageorgiou, 2005; Vermunt and Magidson, 2002). In this probabilistic point of view, every cluster has a different underlying probability distribution from which its data elements are generated. When the distribution functions are known, the problem of finding the clusters reduces to a parameter estimation problem. Given the data elements $Y_1, \dots, Y_n$, each described by a set of features $(y_1, \dots, y_m)$, the prior probability P(z) for cluster C_z with z = 1 , … , K and the conditional multivariate probability density $p(Y|C_z, \theta_z)$, where θ_z is the unknown parameter vector, the mixture probability density for the whole data set can be expressed as

$$p(Y|\theta) = \sum_{z=1}^{K} P(C_z) p(Y|C_z, \theta_z) \quad (1)$$

Following the maximum likelihood approach, the unknown parameter vector is often estimated by means of the expectation-maximization algorithm. Outliers are handled by adding one or more classes, representing a different multivariate distribution for outliers (Fraley and Raftery, 2002). Typically if a small cluster appears which is hard to profile by means of the cluster-dependent distributions, one has found a group of outliers. Other possibilities to handle outliers are by using exploratory analysis in advance (Moustaki and Papageorgiou, 2005).

Finite mixture models have been implemented in different software packages, such as MCLUST, GMDD, AutoClass, Multimix, EMMX, SNOB (Xu and Wunsch, 2005) and Latent Gold. All these software packages use the finite mixture model expressed by equation 1, but differ in regard to the implemented algorithm and probability distributions for $p(Y|C_z, \theta_z)$.

For this research, we selected the model-based clustering technique implemented by the software package Latent Gold. The remainder of this section will focus on the model-based clustering algorithm as implemented by this particular piece of software. For a more complete discussion and review of existing clustering techniques, the interested reader is referred to some classical texts as Gordon (1999) or Kaufman and Rousseeuw (2005) and the elaborated survey paper by Xu and Wunsch (2005). It should be noted that the goal of this paper is to investigate whether clustering techniques can be used to reduce traffic accident heterogeneity. Our choice for LCC is merely based on advantages held by this technique which will be discussed infra and which makes it an appropriate clustering method for this type of research.

Latent Gold implements the aforementioned finite mixture model, but allows the user to make the assumption that the variables y_j, describing the data elements Y_i are independent within each clusters. This assumption greatly reduces the parametric complexity of the model and allows mixing descriptive variables y_j of different types (e.g. categorical and continuous) easily. We can now rewrite the mixture probability density for the whole data set as

$$p(Y|\theta) = \sum_{z=1}^{K}\left[P(C_z)\prod_{j=1}^{m} p\left(y_j \middle| C_z, \theta_z\right)\right] \quad (2)$$

Different probability density functions can be used in Latent Gold. If the descriptive feature is continuous, a normal Gaussian distribution is used, if the descriptive variable is nominal, a multinomial distribution is selected, for an ordinal variable, an adjacent-category ordinal logistic regression model is applied, while the Poisson distribution is used for count variables (Vermunt and Magidson, 2005). After estimation of the parameter vector θ, the underlying statistical model assigns a set of posterior probabilities p_{ik} of belonging to cluster C_k to each data element Y_i. This implies that a traffic accident can belong to a cluster of traffic accidents with pedestrians involved with 40% probability and to a cluster of traffic accidents on a crossroad with 60% probability. In this regard, LCC resembles fuzzy clustering (Höppner et al., 1999). The estimated underlying statistical model also allows calculating the cluster probabilities for new cases.

Furthermore, compared with traditional distance-based clustering techniques, LCC holds other advantages which makes it appropriate for clustering traffic accident data. Firstly, contrary to techniques such as K-means or hierarchical clustering algorithms, no distance measure has to be selected and normalization of the data doesn't have any effect on the final

clusters (Vermunt and Magidson, 2002; Hair et al., 1998). Secondly, in contrast to hierarchical clustering techniques, LCC doesn't have large memory demands, allowing it to build models from large datasets (Hair et al., 1998, Brijs, 2002). Thirdly, unlike K-means algorithms, the LCC technique provides several statistical criteria to choose the number of clusters, such as the information criteria BIC, AIC and CAIC, which take the model's parsimony into account and are based on the model's likelihood (Vermunt and Magidson, 2002). Finally , LCC allows the researcher to easily include variables of different scales (nominal, ordinal, count, continuous) (Vermunt and Magidson, 2002, Brijs, 2002).

3. Data

Belgium is a small and mainly urbanized country with 10.3 million inhabitants on 30 528km². Because large disparities exist in terms of urbanization and population distribution within the country (Mérenne et al., 1997), our analysis was limited to the Brussels Capital region. This administrative region, which comprises nineteen cities and municipalities, among which the Belgian capital city, covers an area of 161.4km² and counts almost one million inhabitants. Furthermore, the Brussels Capital region covers 1 881 kilometers road which accounted for 3.18 billion vehicle kilometers traveled in 2005 (FPS Economy – Directorate-general Statistics Belgium; FEBIAC).

In Belgium, every traffic accident with casualties is officially registered by a police officer at the traffic accident site by means of the "Analysis Form for Traffic Accidents with casualties". These data are collected, digitized and made public by the Directorate-general Statistics Belgium. From these data, we retrieved all two road users traffic accidents for the period between 1997 and 1999. This period should be short enough to embank structural changes in road and traffic conditions, but still long enough to limit any biased effects for random fluctuations.

The final dataset contains 29 variables, describing accident, vehicle, road user and environmental related aspects of the traffic accident. Table 1 gives an overview of all traffic accident variables. Some of these variables were measured at the traffic accident level, while others were measured at the road user level (both first and second). Traffic accidents with pedestrian or motorcyclists/bicyclists involved were recoded, such that the pedestrian or motorcyclists/bicyclists were always coded as the first road user. In total, the final dataset contains 4028 traffic accidents.

4. Results

4.1 Cluster analysis

In this study, we entered all variables mentioned in table 1 into the cluster analysis, except for the variables *consequence* and *detrimcounts*. These were excluded because LCC creates clusters such that all variables are independent from each other within each cluster. However, we do not want *consequence* or *detrimcount* to be locally independent because they are possible dependent variables in subsequent injury risk analysis.

The BIC, AIC and CAIC were used to choose the number of clusters in the final model. These statistical figures measure the model fit and simultaneously correct for the model's complexity (a more parsimonious model is better). Based on figure 1 (a lower score is better), the seven cluster model was selected as our final model. From seven clusters on, BIC and CAIC hardly show any additional improvement. Furthermore, in order to assess the quality of the clustering solution we calculated the entropy criterion (McLachlan and Peel, 2000) as in equation (3) where p_{ik} denotes the posterior probability that case *i* belongs to cluster *k* and with the convention that $p_{ik} \ln(p_{ik}) = 0$ if $p_{ik} = 0$.

$$I(k) = 1 - \frac{\sum_{i=1}^{n}\sum_{z=1}^{k} p_{ik}\ln(p_{ik})}{n\ln(1/k)} \qquad (3)$$

In case of perfect classification the criterion equals to 1 and for the worst case clustering the value of the criterion is 0. For our data, we found that *I(7) = 0.92,* which indicates a very good separation between the clusters.

The seven cluster model provides cluster-dependent univariate distributions for each variable which allow us to identify each cluster as a specific traffic accident type. In order to describe each cluster concisely, we focus on the skewed feature distributions which differ between the clusters. For example, if one cluster contains 99% weekend accidents, while the other clusters have more balanced distributions for the feature “weekend”, one can identify this cluster as the “weekend traffic accident” cluster. It should be noted that this analysis merely tries to identify various traffic accident types within our data which can be overlapping. Assignment of the traffic accidents to the various clusters are based on the cluster probabilities. As a result, some clusters are described by means of road user characteristics, while other traffic accident types are described by means of road characteristics. The final set of features used for profiling the seven clusters found, is shown in table 2. Additional results are available from the author.

In cluster 1, 95% of all traffic accidents occur on crossroads and more specifically 74% of all traffic accidents occur on crossroads with no traffic lights and no priority roads. We will refer to this cluster as “traffic accidents on crossroads with no traffic lights and no priority roads”.

Furthermore, table 2 reveals that 99% of all accidents within cluster 2 concern a collision with a pedestrian, who is older than 18 in 94% of all cases. We describe this cluster as the “traffic accidents with adult pedestrians”.

Cluster 3 overlaps with cluster 1 because almost all traffic accidents occur on a crossroad (99%). However, compared with cluster 1, the accidents of cluster 3 predominantly occur on crossroads with traffic lights (47% versus 9% for cluster 1). We refer to this cluster as “traffic accidents on crossroads with predominantly traffic lights”.

The best way to discriminate cluster 4 from the other clusters is by means of the second road user’s behavior. In 79% of all traffic accidents within cluster 4, the second road user was not moving. Furthermore, this cluster differs from cluster 5 by the fact that the vehicle of the first road user was predominantly a car. We refer to this cluster as “traffic accidents between a car and a non-moving second road user”.

The fifth cluster discriminates itself from other clusters by means of the vehicle type of the first road user, which is a motorcycle or bicycle in 90% of all cases. We conclude that these are “traffic accidents with a motorcycle or bicycle”.

Cluster 6, contains traffic accidents with pedestrians (99%) who are predominantly younger than 19 (96%). We will refer to this cluster as “traffic accidents with non-adult pedestrians”. Finally, cluster 7 can be described as the cluster with “traffic accidents on highways, national, regional or provincial roads” (99%).

An overview of all clusters and their sizes is given in table 3. Our results seem to confirm Yau’s suggestion that traffic accident data should be segmented on vehicle type (Yau, 2004). Our cluster analysis makes a distinction between traffic accidents with cars, motorcycles or bicycles, or with pedestrians involved. But our analysis also shows that other type of features can be used to segment data, such as type of crossroad or road type.

Finally, it should be noted that the descriptive cluster analysis in this study focuses on finding a concise description for each traffic accident type, which can be useful during the interpretation of our subsequent analyses. However, the results of our cluster analysis also

contain other useful information which can provide interesting insights into the various traffic accident types, which are lost when reducing each cluster to a short one-sentence description.

For example, it seems that traffic accidents with motorcyclists or bicyclists largely occur among people younger than 19, i.e. 42% of all traffic accidents in cluster 5 concerns a non-adult motorcyclist or bicyclist, which is relatively high compared to most other clusters. Secondly, when comparing traffic accidents with adult versus non-adult pedestrians, it seems that only 13% of the adult pedestrians were hidden from the second road user compared to 49% in case of non-adult pedestrians. This indicates that the danger of hidden pedestrians mainly relates to young pedestrians, which can be explained by their height and their inexperience in identifying dangerous situations.

4.2 Injury analysis

4.2.1 Statistical approach

The previous section showed that LCC succeeds in identifying different traffic accident types with good separation and allows us to describe them easily. However, this is only useful if the detected clusters reduce heterogeneity in a way that new information and insights are gained when performing explanatory traffic accident analysis. To test this, we will perform traffic injury analysis for each traffic accident. We will apply the multinomial logit model, which is the most widely applied discrete-outcome modeling approach for accident-severity analysis (Zhang et al., 2000; Bédard et al., 2002; Al-Ghamdi, 2002; Islam and Mannering, 2006; Ulfarsson and Mannering, 2004; Kim et al., 2007, Carson and Mannering, 2001; Shankhar and Mannering, 1996; Valent et al., 2002; Yau, 2004). Our injury risk analysis mainly follows the work of Ulfarsson and Mannering (2004), Islam and Mannering (2001) and Kim et al. (2007).

This approach models the injury severity as,

$$S_{in} = \beta_i X_{in} + \varepsilon_{in} \qquad (4)$$

where S_{in} is the function that determines the probability of discrete outcome i for accident observation n, X_{in} is the vector of accident features that determine the injury severity for accident n, β_i is a vector of estimable coefficients, and ε_{in} is an error term accounting for unobserved effects influencing the injury severity of accident n. It can be shown that if ε_{in} are assumed to be extreme value distributed (cf. McFadden, 1981), then a standard multinomial logit model results.

$$P_n(i) = \frac{EXP[\beta_i X_{in}]}{\sum_{\forall I} EXP[\beta_I X_{In}]} \qquad (5)$$

$P_n(i)$ is the probability that accident n will result in injury outcome i and I is the set of possible accident-severity outcomes.

In addition to the model's coefficient vector β_i, we also calculate the average direct pseudo-elasticity, which captures the percentage change in probability when the k^{th} accident feature is changed from 0 to 1. The elasticity is necessary to correctly judge the relative impact of an explanatory variable, because the sign of a coefficient does not always equal the sign on the change in probability (Greene, 1997). The direct pseudo-elasticity $E_{x_{nk}}^{P_{ni}}$, of the k^{th} feature x_{nk} from the vector x_n, with respect to the probability $P_n(i)$, is computed as

$$E_{x_{nk}}^{P_{ni}} = \frac{P_n\left[\mathrm{i}\middle|x_{nk}=1\right] - P_n\left[\mathrm{i}\middle|x_{nk}=0\right]}{P_{ni}\left[\mathrm{i}\middle|x_{nk}=0\right]} \quad \textbf{(6)}$$

Although the accident injury outcome is an ordinal variable, a multinomial logit model was preferred over an ordered logit or probit model. Many researchers apply a multinomial logit model for this type of research because MNL models allow a variable to have a convex (or concave) effect which pushes away from (towards) the middle injury severity levels and towards (away from) the high and low injury severity levels (Kim et al., 2007; Ulfarsson and Mannering, 2004; Islam and Mannering, 2006; Savolainen and Mannering, 2007).

A problem with the estimates of the MNL model occurs when the independence of irrelevant alternatives (IIA) is violated, i.e. the outcomes share unobserved effects. A possible remedy for this problem, which received attention in traffic research recently, is the nested multinomial logit model, which nests the outcomes that share unobserved effects (Savolainen and Mannering, 2007; Lee and Mannering, 2002; Chang and Mannering, 1999). However, Saccomanno et al. (1996) tested the reliability of different nesting structures for injury severity models and found little difference in predictive power. Whenever the Small and Hsiao (1985) test, which verifies the IIA assumption, indicated that a nested multinomial logit model might be more appropriate, we built all possible nested models and verified if the structure was warranted with the data available (Washington et al., 2003; Savolainen and Mannering, 2007). For all clusters, the ordinary multinomial logit model was the preferred model.

Since the estimated models are conditional on accident occurrence, they do not have an accident-risk interpretation. Rather, the models show which explanatory factors are associated

with increasing probability of particular injury severity categories given that an accident occurred. This approach, which is often applied in injury outcome analysis, avoids the need to know or measure exposure (Kim et al., 2007; Ulfarsson and Mannering, 2004; Lee and Mannering, 2002; Savolainen and Mannering, 2007).

To verify the quality of the models, we calculated ρ^2 which compares the log-likelihood of a model with only constants to the log-likelihood at convergence for the full specification. Furthermore, as recommended by Cox (1961,1962), we tested whether the union of all cluster models differ significantly from the full data model by calculating the test statistic

$$X^2 = -2\left[LL(\beta_F) - \sum_{z=1}^{k} LL(\beta_z) \right] \quad (7)$$

which is χ^2 distributed with *J* degrees of freedom, where $J = \sum_{z=1}^{k} K_z - K_f$ (K_z and K_f are the number of coefficients in the model for respectively cluster C_z and the full data model) and $LL(\beta_z)$ and $LL(\beta_F)$ are the log-likelihoods at convergence for the cluster models and the full data model, respectively.

4.2.2 Statistical model

As dependent variable for our injury outcome analysis, we selected the *consequence for the first road user*, which has three categories, i.e. “Killed or Seriously Injured” (KSI), “Slightly Injured” (SI) and “Not injured” (NI). In all models, NI was selected as the base outcome. Originally, sixteen explanatory variables were entered into the model, which were *age, gender, number of passengers* for both road users, *vehicle type* for the first road user (this variable had too many missing values for the second road user) and *season, weekend, hour,*

road sort, road type, road surface, crossroad, speed limit, accident type. To prevent zero-cell problems in the MNL regression, some of the explanatory variables had to be recoded into a smaller number of categories. Grouping categories together is an acceptable solution for the zero-cell problem if these categories make sense together (Menard, 2001). Furthermore, all explanatory variables are categorical and were recoded into dummy variables. For all sets of dummy variables, one dummy variable was removed from the model to avoid perfect colinearity. In total, eight models were built, one for the full data set and one for each cluster. Traffic accidents were assigned to a cluster if the posterior probability for that cluster was at least 90%. We selected a probability of 90%, which is high enough to have a correct assignment but does not result in too small sample sizes.

In correspondence with Kim et al. (2007), we restricted coefficients to zero which were originally not found significantly different from zero at the 90% level as indicated by an asymptotic t-test. Furthermore, we tested differences between coefficients on one variable across injury severity categories. When there were no statistically significant differences at a 95% level, as indicated by a likelihood ratio test, the coefficients were constrained to be equal. This is done to avoid keeping artificial accuracy in the model.

4.2.3 Results

Rather than performing a complete injury risk analysis for the Brussels Capital region, the objective of these MNL models is to evaluate the usefulness of cluster based traffic accident segmentation. Therefore, we shall focus the discussion on the differences between the various models and are particularly interested to see if the cluster models reveal new information. The estimation results of the eight models can be found in table 4. To conserve space, coefficients statistically insignificant at a 10% confidence level were omitted from the table. All eight models were statistically significant at a 0.1% confidence level, except for clusters 2, 5 and 6 which were still respectively statistically significant at a 6%, 7% and 2% confidence level.

The ρ^2 of the models vary between 3% and 14%, which are reasonable values given the amount of variation in the data. The test if all cluster models together differ from the full data model was significant at 0.1%.

Turning to table 5, which presents the corresponding elasticities, we will study the added value of a clustering analysis preceding the injury-outcome analysis. Firstly, these results confirm the assumption that performing traffic accident analysis on a large heterogeneous data set can obscure significant relations, which was also found by other researchers as mentioned in the introduction. For example, a normal road surface (as opposed to a wet or snowy) was not statistically significant in the full data model. However, according to table 5, a normal road surface lowers the probability of getting slightly injured in a traffic accident on a crossroad with traffic lights with 19%. Furthermore, the average probability of getting killed or seriously injured in a traffic accident on a crossroad with no priority road and no traffic lights decreases with 90% if the first road user drives a car, but increases with 131% and 210% if the first road users drives respectively a motorcycle or a bicycle. None of these results were found with the full data model.

Secondly, the cluster models reveal variation of a variable's effect on the injury outcome probability between different traffic accident types. Table 5 shows that according to the full data model, the probability for the first road user of getting slightly injured in a traffic accident in which the second road user has no passengers increases with 65%. However, the cluster models show that this probability increases with 64% in a traffic accident on a crossroad with no priority road and no traffic lights, with only 54% in a traffic accident on a crossroad with traffic lights, but with no less than 170% in a traffic accident with an adult pedestrian. Furthermore, when the first road user rides a bicycle, the probability of getting slightly injured increases with 210% if he is involved in a traffic accident on a crossroad with no priority road and no traffic lights, but only with 91% if he is involved in a traffic accident

on a crossroad with traffic lights. Note that the full data model did not only hide this difference, it hid this variable completely. In both examples, the cluster-based models reveal a more complete interpretation.

Thirdly, our cluster based models even reveals a different direction of the effect for some features in one or two clusters. While the full data model suggests that being a female first road user increases the probability of getting slightly injured when involved in a traffic accident, cluster models 5 and 6 show that this probability decreases for traffic accidents with motorcycles and bicycles or for traffic accidents with a non-adult pedestrian. Furthermore, the cluster models also show that if a traffic accident with an adult pedestrian occurs on a national, regional or provincial road, the probability of the pedestrian getting slightly injured increases with 74%. On the other hand, if a traffic accident on a crossroad with traffic lights occurs on a national, regional or provincial road, the probability of the first road user getting slightly injured decreases with 23%.

5. Limitations and directions for future research

There are several considerations in order in interpreting the results of our study. First, with regard to the clustering technique, finite mixture modeling or latent class clustering is based on the maximization of a likelihood function. Because the likelihood function is not guaranteed to be concave, the found solution might be a local rather than the global maximum. In this regard, the found solution is dependent on the initial parameter values. To prevent ending up with a local solution, the Latent GOLD program uses 10 sets of random start values (Vermunt and Magidson, 2005). A second consideration relates to the finite mixture model in this study. Our cluster model assumes local independence between the traffic accident variables. This restriction on the cluster-specific covariance matrix can be removed from the model, but comes at a high cost of parametric complexity and computing time.

Future research can study the impact of this limitation on this traffic injury study for the Brussels Capital region. One can for example allow for class-specific covariance between some variables, model-specific covariance between all variables or even class-specific covariance between all variables. One could also build models with different software, allowing for different distributions. For ordinal variables, one could use the cumulative logit model which has the advantage of having effect estimates which are approximate invariant to the choice and number of response categories (Agresti, 2002).

On a more conceptual level, future research could also focus on finding typical groupings in the data, i.e. to identify clusters which can be transferred to other data sets. In such research, it is important to validate the cluster results to provide a degree of confidence (Xu and Wunsch, 2005). One could even try to find a generic list of traffic accident types which can be used to reduce heterogeneity in general, although it is not certain such list exists. One could also try to develop quality checks to validate the cluster descriptions. Thirdly, other ways to dissect the data are possible. For example, the posterior probabilities could be used as weights in subsequent analysis, although one has to be careful with the correct interpretation of these weights.

6. Conclusions

In this paper we examined whether cluster analysis could be used as a traffic accident segmentation technique. We selected latent class clustering as the applied cluster analysis and found that it succeeds in finding various clusters in a heterogeneous traffic accident data set. Furthermore, our research indicated that the traffic accident types, identified by the seven clusters, make sense and add value to subsequent injury analyses. In agreement with the work of Yau (2004), the cluster analysis uses vehicle type as a basis for the segmentation. However, it also finds less trivial variables to segment the data, such as road type and age. The injury analysis performed for each traffic accident type or cluster clearly illustrates the added value

of the cluster-based segmentation. Firstly, the cluster models reveal new variables influencing the injury outcome. Secondly, the results illustrate that cluster models can sometimes provide a more complete interpretation of the effect of an independent variable on the injury outcome, by showing different magnitudes among different traffic accident types. Thirdly, the cluster models also reveal that the effect of a single independent variable can differ in direction between different traffic accidents. These results confirm the existence of three types of information which can remain hidden due to data heterogeneity, which was already illustrated by the results of other authors.

Acknowledgements

The authors would like to thank P. Savolainen for his advice and feedback on building the nested logit models. They also acknowledge extensive comments from the reviewers.

References

Agresti, A., 2002. Categorical Data Analysis. Wiley series in probability and statistics.

Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Anal. Prev. 34 (6), 729-741.

Arabie, P., Hubert, L.J., 1994. Cluster Analysis in Marketing Research. In Advanced Methods of Marketing Research, R.P. Bagozzi ed. Oxford: Blackwell, 160-189.

Bédard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Anal. Prev. 34 (6), 717-727.

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

Bock, H., 1996. Probabilistic models in cluster analysis. Computational Statistics and Data Analysis 23 (1), 5-28.

Brijs, T., 2002. Retail Market Basket Analysis. A Quantitative modeling approach. PhD Dissertation. University of Hasselt, Diepenbeek, Belgium.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. Accident Anal. Prev 33 (1), 99-109.

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Anal. Prev. 38 (6), 1019-1027.

Chang, L.-Y., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Anal. Prev. 31 (5), 579-592.

Cox, D.R., 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkely Symposium on Mathematics, Statistics and Probability. University of California Press, Berkeley.

Cox, D.R., 1962. Further results on tests of separate families of hypotheses. J. R. Stat. Soc. B24, 406-424.

Everitt B., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold.

FEBIAC (Belgian Federation of the Car and Two-wheeler Industries [online]. Available from World Wide Web: (http://www.febiac.be).

FPS Economy – Directorate-general Statistics Belgium [online]. Available from World Wide Web: (http://statbel.fgov.be).

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

Fraley, C., Raftery, A.E., 2002. Model-based clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97 (458), 611-631.

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

Geurts, K., Thomas, I., Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. Accident Anal. Prev. 37 (4), 787-799.

Gordon, A., 1999. Classification. Chapman & Hall.

Green, P.E., 2004. Practice makes perfect. Marketing Research 16 (2), 8-14.

Greene, W.H., 1997. Econometric Analysis. Prentice Hall.

Hair, J.F.Jr., Anderson, R.E., Tatham, R.L., Black, W.C, 1998. Multivariate Data Analysis. Prentice Hall.

Hartigan, J., 1975. Clustering Algorithms. Wiley.

Höppner, F., Klawonn, F., Kruse, R., 1999. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition. Wiley.

Hosking, J., Pednault, E., Sudan, M., 1997. A statistical perspective on data mining. Future Gen. Comput. Syst. 13(2-3), 117-134.

Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. Accident Anal. Prev. 37(2), 267-276.

Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering 16 (11), 1370-1386.

Kaufman, L., Rousseeuw, P., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.

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

Kim, J.-K., Sungyop, K., Ulfarsson, G.F., Porello, L.A., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accident Anal. Prev. 39 (2), 238-251.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Anal. Prev. 34 (2), 149-161.

McLachlan, G., Peel, D., 2000. Finite mixture models. Wiley series in probability and statistics.

Menard, S., 2001. Applied Logistic Regression Analysis. Second Edition. SAGE Publications.

Mérenne, B., Van der Haegen, H., Van Hecke, E., 1997. Diversité territoriale. Bulletin du Crédit Communal, 202.

Moustaki, I., Papageorgiou, I., 2005. Latent class models for mixed variables with applications in Archaeometry. Computational Statistics and Data Analysis 48 (3), 659-675.

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

Quddus, M., Noland, R., Chin, H., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Safety Res. 33 (4), 445-462.

Saccomanao, F.F., Nassar, S.A., Shortreed, J.H., 1996. Reliability of statistical road accident injury severity models. Transport. Res. Rec. 1542, 14-23.

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Anal. Prev. 39 (5), 955-963.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Safety Res. 27 (3), 183-194.

Small, K.A., Hsiao, C., 1985. Multinomial Logit Specification Tests. Int. Econ. Rev. 26, 619-627.

Ulfarsson, G., Mannering, F., 2004. Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accident Anal. Prev. 36 (2), 135-147.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accident Anal. Prev. 34 (1), 71-84.

Vermunt, J.K., Magidson, J., 2005. Technical Guide for Latent GOLD 4.0: Basic and Advanced. Belmont Massachusetts: Statistical Innovations Inc.

Vermunt, J.K., Magidson, J., 2002. Latent Class Cluster Analysis. In Applied Latent Class Analysis, Hagenaars J.A. and McCutcheon A.L. eds., 89-106.

Washington, S., Karlaftis, M., Mannering, F.L., 2003. Statistical and econometric methods for transportation data analysis. CRC Press.

Xu, R., Wunsch II, D., 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645-678.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Anal. Prev. 36 (3), 333-340.

Zajac, S., Ivan, J., 2003. Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut. Accident Anal. Prev. 35 (3), 369-379.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Anal. Prev. 32 (1), 117-125.

Table 1: Traffic accident variables

Variable	Aspect	Level	Values
Vehicle type	Vehicle	Road user	car; medium sized vehicle; large truck; large bus; motorcycle; bicycle; pedestrian; other
Gender	Road user	Road user	female; male
Age	Road user	Road user	0-15; 16-17; 18-21; 22-29; 30-39; 40-49; 50-59; 60-65; 65+
Crossroad	Environmental	Accident	crossroad with traffic lights; crossroad with priority road; crossroad with right of way or a police officer; no crossroad
Built-up area	Environmental	Accident	inside built-up area; outside built-up area
Road type	Environmental	Accident	highway; national, regional or provincial road; local road
Road sort	Environmental	Accident	single roadway; divided roadway
Speed limit	Environmental	Accident	30km/h; 50km/h; 60km/h; 90km/h; 120km/h
Speed limit diff.	Environmental	Accident	whether or not there is a difference between the speed limits on roads intersecting at a crossroad (yes/no)
Black zone	Environmental	Accident	yes; no
Road structure	Environmental	Accident	roundabout; bridge or viaduct; tunnel; level crossing; sharp bend; school, recreation centre or bus stop; signalization on the road; steep descent
Weekend	Traffic accident	Accident	yes (Monday 2h – Friday 21h); no (Friday 21h – Monday 2h)
Hour	Traffic accident	Accident	morning (7h-9h); early afternoon (10h-12h); afternoon (13h-15h); evening rush (16h-18h); evening (19h-21h); night (22h-6h)
Season	Traffic accident	Accident	winter; spring; summer; fall
Hidden Pedestrian	Traffic accident	Accident	whether or not the pedestrian was visible for the other road user (yes/no)
Missing Safety	Traffic Accident	Accident	whether or not some safety measures were missing, such as wearing a safety belt or helmet (yes/no)
Rain	Traffic Accident	Accident	no rain; rain
Road surface	Traffic Accident	Accident	normal road surface; wet or snowy road surface
Accident type	Traffic Accident	Accident	frontal collision; collision from behind; lateral collision; collision with a pedestrian; collision with an obstacle on the road; collision with an obstacle next to the road
Passenger positions	Traffic Accident	Accident	at least one of the road users had passengers in front and in the back of the vehicle; there were passengers

			involved, but no car had passengers in the front and in the back; there were no passengers in the vehicles of both road users
Detrimcounts	Traffic Accident	Accident	measures the severity of the traffic accident by summing the number of slightly injured, the number of seriously injured multiplied by three and the number of deadly injured multiplied by five
Number passengers	Traffic Accident	Road user	0; 1; 2; 3+
Behavior	Traffic Accident	Road user	ignores red light; fails to give right of way; crosses a full white line; passes incorrectly; makes an evasive maneuver; incorrect location on the road; loss of control; not enough distance kept; fall; normal behavior
Dynamics	Traffic Accident	Road user	road user travels at constant speed; road user brakes in order to stop; road user accelerates or starts; road user is not moving
Consequence	Traffic Accident	Road user	deceased; seriously injured; slightly injured; no injuries

Figure 1: Evolution of BIC, AIC, CAIC when adding clusters to the model

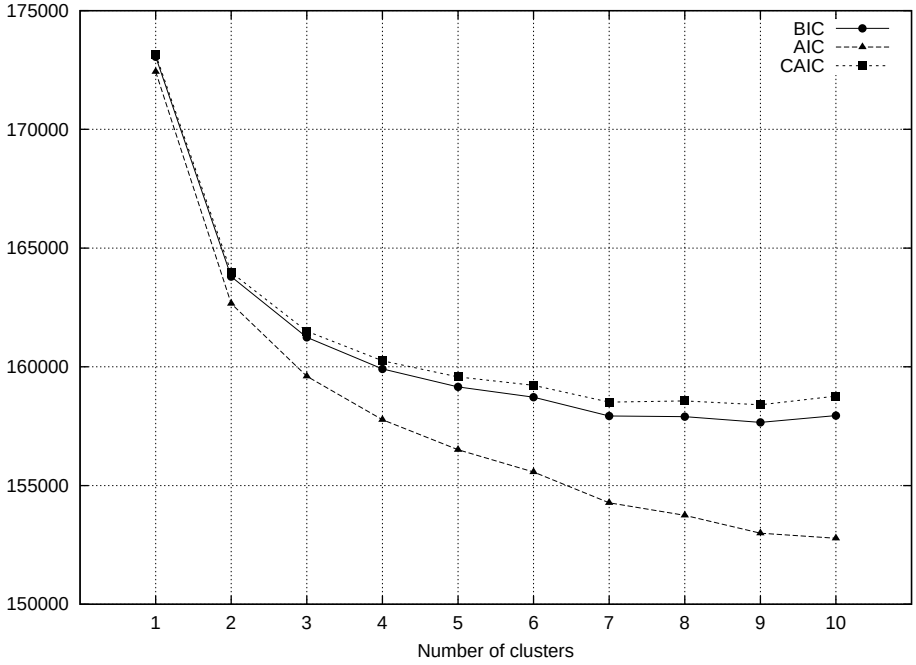

Table 2: Features and their probability in each cluster

Variable – value	**Clu1**	**Clu2**	**Clu3**	**Clu4**	**Clu5**	**Clu6**	**Clu7**
Accident type: *collision with a pedestrian*	0%	99%	0%	0%	0%	99%	6%
Crossroad: *crossroad without traffic lights or priority road*	74%	26%	40%	21%	40%	15%	5%
Crossroad: *no crossroad*	5%	48%	1%	56%	33%	76%	42%
Built-up area: *Outside built-up area*	1%	1%	1%	1%	1%	0%	47%
Road type: *Highway, national, regional or provincial road*	25%	25%	55%	23%	24%	7%	99%
Age road user 1: *0-18 years old*	14%	11%	16%	16%	42%	96%	11%
Dynamics road user 2: *Road user is not moving*	0%	11%	0%	79%	74%	35%	0%
Vehicle type road user 1: *Motorcycle or bicycle*	8%	0%	12%	17%	90%	0%	7%
Vehicle type road user 1: *Car*	85%	0%	81%	76%	9%	0%	82%

Table 3: Traffic accident types

Cluster	Traffic accident type	Size
1	traffic accidents on crossroads with no traffic lights and no priority road	23%
2	traffic accidents with an adult pedestrian	19%
3	traffic accidents on crossroads with predominantly traffic lights	15%
4	traffic accidents between a car and a non-moving second road user	15%
5	traffic accidents with a motorcycle or bicycle	14%
6	traffic accidents with a non-adult pedestrians	10%
7	traffic accidents on highways, national, regional or provincial roads	4%

Table 4: Multinomial logit estimation results for first road user crash-injury severity

Model:	F		1		2		3		4		5		6		7	
Variable	Coef.	*S.E..*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
age road user 1: <30 [KSI]	-0.61	*0.27*														
age road user 2: <30 [KSI]	1.77	*1.03*														
age road user 2: 30-60 [KSI]	1.85	*1.02*														
collision from behind [KSI]									-1.93	*1.17*						
collision with pedestrian [KSI]	-0.57	*0.30*														
collision with obstacle [KSI]	0.53	*0.25*	1.04	*0.45*					2.33	*1.06*						
crossroad with a priority road [KSI]	-1.32	*0.72*														
crossroad with no priority road and no traffic lights [KSI]	-0.27	*0.08*														
afternoon [KSI]	-0.29	*0.11*														
road user 2 had no passengers [KSI]									0.71	*0.27*						
winter [KSI]			-1.41	*0.82*												
spring [KSI]	-0.71	*0.35*	-2.52	*1.09*												
road user 1 drove a car [KSI]			-2.35	*0.61*												
road user 1 drove a motorcycle [KSI]			1.77	*0.37*												
road user 1 drove a bicycle [KSI]			3.41	*1.05*												
highway [KSI]					2.44	*1.16*										
constant [KSI]	-4.43	*1.02*	-1.85	*0.66*	-4.08	*0.51*	-3.38	*0.59*	-4.26	*0.99*	-4.80	*0.71*	-4.92	*0.71*	-2.71	*0.46*
age road user 1: 30-60 [SI]			-0.35	*0.16*												
female road user 1 [SI]	0.61	*0.10*	1.08	*0.19*			1.21	*0.30*	0.71	*0.27*	-1.26	*0.75*	-2.40	*1.04*	1.86	*0.45*
female road user 2 [SI]	-0.45	*0.10*	-0.88	*0.19*			-0.79	*0.28*	-0.80	*0.27*	0.57	*0.34*			-0.86	*0.41*
collision from pedestrian [SI]	-1.50	*0.19*													-1.39	*0.85*
collision with obstacle [SI]	0.53	*0.25*	1.04	*0.45*												
crossroad with no priority road and no traffic lights [SI]	-0.27	*0.08*														
afternoon [SI]	-0.29	*0.11*														
evening [SI]	-0.17	*0.09*							-0.72	*0.30*						
road user 2 had no passengers [SI]	0.68	*0.10*	0.73	*0.18*	1.06	*0.63*			0.71	*0.27*						
winter [SI]	0.17	*0.10*														
summer [SI]	0.27	*0.10*			0.66	*0.37*										
road user 1 drove a car [SI]	0.97	*0.15*					0.94	*0.39*	1.46	*0.52*						
road user 1 drove a motorcycle [SI]	0.47	*0.17*	1.77	*0.37*					2.69	*0.79*						
road user 1 drove a bicycle [SI]			3.41	*1.04*			1.52	*0.93*								

national, regional or provincial road [SI]					0.61	*0.35*	-0.49	*0.25*								
normal roadsurface [SI]							-0.40	*0.24*								
constant [SI]	-1.66	*0.18*	-1.06	*0.18*	-3.24	*0.68*			-1.81	*0.57*	-2.23	*1.14*	-1.59	*0.57*		
Number of observations	3282		799		535		376		333		296		297		142	
Log Likelihood at zero	-2529.18		-675.92		-214.53		-318.25		-303.16		-151.19		-91.56		-129.18	
Log Likelihood at convergence	-2006.40		-507.27		-202.35		-265.81		-236.37		-140.91		-78.70		-102.80	
ρ^2	0.12		0.14		0.03		0.09		0.12		0.04		0.08		0.11	

Coefficients that were not significant at the 90% level were omitted from the table. “No injury” is the base outcome with coefficients restricted at zero. “F” represents the full data model and models 1 until 7 respectively represent cluster model 1 until 7.

Table 5: Elasticities in percent

Model:	F	1	2	3	4	5	6	7
Variables								
KSI Elasticities								
age road user 1: <30	-45%							
age road user 2: <30	455%							
age road user 2: 30-60	506%							
collision from behind					-85%			
collision with pedestrian	-21%							
collision with obstacle	41%	74%			750%			
crossroad with a priority road	-73%							
crossroad with no priority road and no traffic lights	-17%							
afternoon	-18%							
road user 2 had no passengers					54%			
winter		-75%						
spring	-50%	-92%						
road user 1 drove a car		-90%						
road user 1 drove a motorcycle		131%						
road user 1 drove a bicycle		210%						
highway			846%					
SI Elasticities								
age road user 1: 30-60		-20%						
female road user 1	52%	88%		79%	51%	-67%	-90%	144%
female road user 2	-28%	-44%		-37%	-39%	59%		-39%
collision from pedestrian	-68%							-60%
collision with obstacle	41%	74%						
crossroad with no priority road and no traffic lights	-17%							
afternoon	-18%							
evening	-12%				-36%			
road user 2 had no passengers	65%	64%	170%		54%			
winter	13%							
summer	21%		82%					
road user 1 drove a car	104%			83%	185%			
road user 1 drove a motorcycle	38%	131%			186%			
road user 1 drove a bicycle		210%		91%				
national, regional or provincial road			74%	-23%				
normal roadsurface				-19%				